Biotic Prediction

Building the Computational Technology Infrastructure
for Public Health and Environmental Forecasting

Baseline Software Design Document

BP-BSD-1.2

Task Agreement: GSFC-CT-1

October 17, 2002

Milestone E: Code Baseline Completed, Due 07/15/02

Document performance characteristics and time/space complexity of existing PlantDiversity code and modeling process for two canonical examples: Rocky Mountain National Park (RMNP) and the Cerro Grande Fire Site (CGFS). Determine appropriate multipliers, m and n, to be used in Milestones F and G respectively. Deliver initial version of Requirements and Software Design Documents. Documented source code made publicly available via the Web.

The following documentation shall be provided in fulfillment of these milestones:

- Title of the agreement and agreement number.
- Text of the milestone and its due date.
- A written description of the problem being solved to demonstrate the required improvement.
- A written description of the computer code(s) used to meet the milestone, including descriptions of the algorithms, numerical methods, and parallel implementation.
- If the code is a parallel code, a scaling analysis showing the performance of the code on several numbers of processors including the number used to meet the milestones.
- Documentation as identified in the appropriate milestone.
- The location of an FTP or Web site where NASA may obtain a copy of the computer code(s) in source language form, and any test datasets, makefiles, or other information necessary for NASA to independently verify the achievement of the milestone. This data may also be made available to NASA by noting its location on the file system of a computing system where it can be run. If this system is not a computing system provided by the CT Project, provision must be made for access by CT staff to perform the validation.
- A summary of the scientific or computational significance of achieving the milestone, including graphics if appropriate.

Contents

List of Tables

List of Figures

1 Overview

1.1 Introduction

This project will develop the high-performance, computational technology infrastructure needed to analyze the past, present, and future geospatial distributions of living components of Earth environments. This involves moving a suite of key predictive, geostatistical biological models into a scalable, cost-effective cluster computing framework; collecting and integrating diverse Earth observational datasets for input into these models; and deploying this functionality as a Web-based service. The resulting infrastructure will be used in the ecological analysis and prediction of exotic species invasions. This new capability will be deployed at the USGS Midcontinent Ecological Science Center and extended to other scientific communities through the USGS National Biological Information Infrastructure program.

1.2 Referenced Documents

Table 1. Referenced Documents

Document Title	Version	Date
Software Engineering / Development Plan	1.0	2002-04-08
Concept of Operations	1.6	2002-10-17
Software Requirements Document	1.2	2002-10-17

1.3 Document Overview

This document, the *Baseline Software Design Document*, describes the software design and architecture for the Baseline Software. This represents the heritage software that is being transformed and incorporated into a new software system, the *Invasive Species Forecasting System* (ISFS).

Section 2 discusses the problem class that the system addresses, and the approach that the system uses to address that problem.

Section 3 describes the specific model used in the baseline software, *PlantDiversity*, along with an introduction to the numerical methods it uses.

Section 4 describes the practical implementation of the modeling process into a software processing flow.

Section 5 details the use of the software as applied to the canonical cases. This includes a discussion of the performance characteristics of the model for the cases.

Our plans for improving the performance of the core model are discussed in section 6.

Some information about the delivery of the software are in section 7.

2 Problem Class

USGS has implemented a heritage modeling process, which we refer to as *PlantDiversity*, that we are transforming into a coherent *Invasive Species Forecasting System* (ISFS). The ISFS will be used to analyze the past, present, and future geospatial distributions of living components of Earth environments.

2.1 Introduction

Many of the most important science questions we hope to address by modeling the Earth system involve understanding where a particular species or group of organisms exist at a given time. For example, in order to understand the effects of land cover and land use change, we may wish to know historically whether certain types of plants or animals were once present in a region of interest. In real-time, we may wish to know the public health risk for vector-mediated diseases, such as Hantavirus Pulmonary Syndrome or Lyme Disease. In these cases, it would be important to know the current distributions of deer mice and black-legged ticks, which are responsible, respectively, for transmitting these diseases to humans. In order to understand how the health and functioning of entire ecosystems are being influenced by the invasion of exotic species, we may wish to predict the future distribution patterns of key native and non-native organisms.

Determining the geospatial distribution of living things across various time frames and time scales requires an understanding of the natural history of the organisms in question. It involves the formulation of sometimes complex theory about their behavior, reproduction, and movements through the environment, and the subsequent reification of these theories into models, simulations, and computational analyses. It draws upon diverse and heterogeneous data, including remotely sensed data, ground-based point data, and data about past life from natural history collections. Increasingly, there is a need for interactive visualization of results and the ability to fold results into decision support systems and other mechanisms that enable the development of effective policy and action. From both a scientific and technological perspective, these are nontrivial problems.

Our overarching goal is to enable the ecological, environmental, and public health communities by expanding their participation in high-performance computing. We propose to start the development of a generalized computational technology infrastructure for these communities by focusing on a class of landscape-scale geostatistical models that predict the distributions of living organisms. We will work specifically with a well-understood ecological model, *PlantDiversity*. The *PlantDiversity* model is currently being used to perform landscape-scale assessments of plant diversity and to predict exotic plant invasions in US parks and wilderness areas. This is an important modeling process, and it represents an important class of codes. By working on the *PlantDiversity* model, we can characterize common elements and identify functionality that might be abstracted away from the core model and delivered as general Web-based services to the broader scientific community.

2.2 Objectives

The specific objectives of this work include the following:

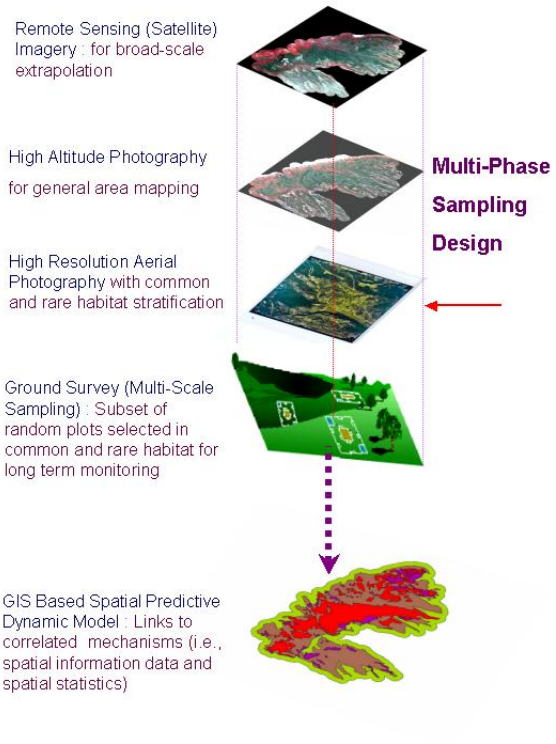
- Create a high-performance, parallel implementation of the *PlantDiversity* (invasive species) model code;
- Document the use of software engineering techniques that foster reproducibility and community-wide software process improvements in these domains;

- Engage an extended community of scientists through the established NBII community infrastructure program; and
- Empower the ecological, environmental, and public health communities by expanding their participation in high-performance computing and greater use of NASA data.

2.3 Approach

Predictive spatial models developed from multi-scale data are an excellent example of data synthesis for natural resource management and public health (Chong *et al.*, 2000; Glass, 2000; Kalkhan *et al.*, 2000a, 2000b). Spatial statistics and geostatistics provide a means to develop spatial models that can be used to correlate coarse scale geographic information (e.g., digital elevation models, burn areas, remotely sensed data) with multi-scale field measurements of biotic and abiotic variables (Kalkhan and Stohlgren, 2000). Integral to the creation of spatial models is the collection of appropriate data. Kalkhan *et al.* (1998) and Stohlgren *et al.* (1997a; 1997b; 1997c) have developed a multi-phase, multi-scale sampling approach that involves stratification of areas of interest from remotely sensed data, random location of field sampling points within strata, and sampling with multi-scale plots. Data collection from multi-scale plots allows extrapolation of results to larger scales with calculable error (Figure 1).

The ability to model small-scale variability in landscape characteristics requires the generation of full-coverage maps depicting characteristics measured in the field (Reich *et al.*, 1999). While many spatial datasets describing land characteristics have proven reliable for macro-scale ecological monitoring, these relatively coarse scale data fall short in providing the precision required by more refined ecosystem resource models (Gown *et al.*, 1994). Spatial statistics and geostatistics provide a means to develop spatial models that can be used to correlate coarse scale geographical data with field measurements of biotic variables. This general landscape analysis approach is being used successfully to address a range of natural resource and public health issues, including invasive species (Stohlgren *et al.*, 1998, 1999a,b; Kalkhan and Stohlgren, 2000; Kalkhan *et al.*, 2000a,b,c), detecting “hot spots” of native and exotic plan diveristy and rare/unique habitats (Agee and Johnson, 1988; Noss, 1983; LaRoe, 1993; McNaughton, 1993), detecting habitats vulnerable to invasive and rapid spread of exotic plant species (Stohlgren *et al.*, 1999a), and determining vegetation and soil response to fire (Kalkhan, *et al.*, 2002).



Multi-phase sampling refers to sampling at multiple levels from remotely sensed data (e.g. Landsat TM imagery, aerial photos) to ground data. Multi-scale sampling refers to the use of sampling plots that contain subplots of various sizes. This approach can be applied to biotic and abiotic variables such as exotic species abundance, native species distribution, fire conditions, and other ecological studies. (Adapted and modified from Kalkhan *et al*., 1995, 1998)

Figure 1. Multi-phase Sampling Design

3 Model Description

In this section, we provide a general description of the candidate model that will be the focus of our development efforts. Additional information about the geostatistical methods described here may be found in Isaaks and Srivastava's *Applied Geostatistics* (1989). The *PlantDiversity* model is currently being used to identify areas at risk for exotic plant species invasions at the Cerro Grande Fire Site near Los Alamos, New Mexico (Kalkhan *et al.*, 2002), in Rocky Mountain National Park, Colorado (Chong *et al.*, 2000; Kalkhan *et al.*, 2000a, Kalkhan *et al.*, 2000b), and Grand Staircase-Escalante National Monument, Utah (Kalkhan *et al.*, 2000c).

3.1 Modeling Large-Scale Spatial Variability

As shown in Figure 2, the process begins with stepwise regression and trend surface analysis for geographical variables and measures of focal taxa to evaluate large-scale spatial variability in a study area. The functional form of this model is defined as:

$$\Phi_0 = \sum_{i+}^{p} \sum_{j \leq p}^{p} \beta_{ij} x_{10}^{i} x_{20}^{j} + \sum_{kto} q \gamma_k y_{k0} + \eta_0 \tag{1}$$

where, β_{ij} are the regression coefficients associated with the trend surface component of the model, γ_k are the regression coefficients associated with the q auxiliary variables, y_{k0}, available as a coverage in the GIS data base, and η_0 is the error term which may or may not be spatially correlated with its neighbors (Kallas, 1997; Metzger, 1997).

Stepwise multiple regression analysis is used first to identify the best linear combination of independent variables. It also allows us to explore the variation in predicting total, exotic, and native plant species richness as a function of the TM bands, derived vegetation indices, tasseled cap transformation indices, slope, aspect, and elevation. The selected independent variables are used in an Ordinary Least Square (*OLS*) procedure to describe large-scale variability estimates.

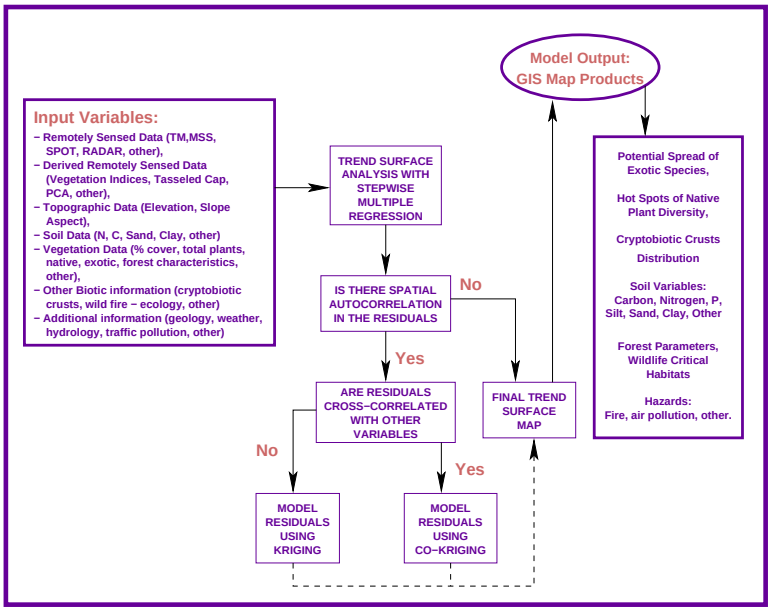
OLS estimators are used to fit the model if the variable of interest has a linear relationship with the geographical coordinates of the sample plots, the digital number (DN) value of any of the Landsat TM bands, and the topographic data. In addition, the least squares method fits a continuous, univariate response as a linear function of the predicted variable. This trend surface model represents continuous first order spatial variation. Akaike's Information Criteria "AIC", (Brockwell and Davis 1991, Akaike 1997) is used as a guide in selecting the number of model parameters to include in the regression model where:

$$\text{AIC} = -2(\text{max log likelihood}) + 2(\text{number of parameters}) \tag{2}$$

When using maximum likelihood as a criterion for selecting between models of different orders, there is the possibility of finding another model with equal or greater likelihood by increasing the number of parameters (Metzger 1997). Therefore, the AIC allows for a penalty for each increase in the number of parameters. Using this criterion, a model with a smaller AIC is considered to have a better fit. While, the model is kept as simplistic as possible, a more complex model could be used if the situation warrants it.

3.2 Modeling Small-Scale Spatial Variability

In the next stage of the model building process the residuals from the trend surface models are analyzed for spatial dependencies. This is accomplished using spatial auto-correlation and cross-correlation



The development of predictive spatial models for exotic plant species, uncertainty, forest parameters, soil variables, and other biotic and abiotic factors relies on the creation of trend surface maps with stepwise multiple regression residuals using and OLS procedure.

Figure 2. Model Flow Chart

statistics. If the residuals are cross-correlated with other variables, we can use co-kriging to interpolate the residuals. However, if the residuals are not cross-correlated, we use ordinary kriging. Finally, the weights associated with the kriging and co-kriging models are estimated as a function of the spatial continuity of the data (Isaaks and Srivastiva 1989). This estimation can be accomplished using a sample variogram to describe spatial continuity. With spatial data, the variation of the samples generally changes with distance. In other words, the variogram is a measure of how the variance changes with distance. The variogram and cross-variogram models used in this analysis were considered "basic" models, meaning they are simple and isotropic (Reich *et al.* 1999). They include, Gaussian, spherical, and exponential models. Since the primary focus of our work in this project involves improving the performance of kriging, we now describe this technique in some detail.

Kriging is a method of interpolation named after a South African mining engineer named D. G. Krige who developed the technique in an attempt to more accurately predict ore reserves. Over the past several decades kriging has become a fundamental tool in the field of geostatistics. Kriging is based on the assumption that the parameter being interpolated can be treated as a regionalized variable. A regionalized variable is intermediate between a truly random variable and a completely deterministic variable in that it varies in a continuous manner from one location to the next and therefore points that are near each other have a certain degree of spatial correlation, but points that are widely separated are statistically independent (Davis, 1986). Kriging is a set of linear regression routines which minimize estimation variance from a predefined covariance model.

The first step in ordinary kriging is to construct a variogram from the scatter point set to be interpolated. A variogram consists of two parts: an experimental variogram and a model variogram. Suppose that the value to be interpolated is referred to as f. The experimental variogram is found by calculating the variance (g) of each point in the set with respect to each of the other points and plotting the variances versus distance (h) between the points. Several formulas can be used to compute the variance, but it is typically computed as one half the difference in f squared.

Once the experimental variogram is computed, the next step is to define a model variogram. A model

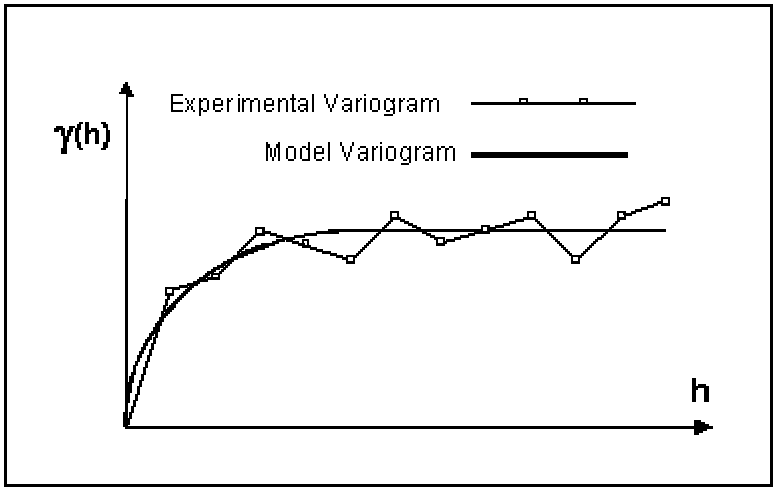


Figure 3. Experimental and model variogram used in kriging

variogram is a simple mathematical function that models the trend in the experimental variogram. As can be seen in the above figure, the shape of the variogram indicates that at small separation distances, the variance in f is small. In other words, points that are close together have similar f values. After a certain level of separation, the variance in the f values becomes somewhat random and the model variogram flattens out to a value corresponding to the average variance.

Once the model variogram is constructed, it is used to compute the weights used in kriging. The basic equation used in ordinary kriging is as follows:

$$F(x,y) = \sum_{i=1}^{n} w_i f_i \tag{3}$$

where n is the number of scatter points in the set, f_i are the values of the scatter points, and w_i are weights assigned to each scatter point. The weights used in kriging are based on the model variogram. For example, to interpolate at a point P based on the surrounding points P_1, P_2, and P_3, the weights w_1, w_2, and w_3 must be found. The weights are found through the solution of the simultaneous equations:

$$\begin{aligned} w_1 S(d_{11}) + w_2 S(d_{12}) + w_3 S(d_{13}) &= S(d_{1p}) \\ w_1 S(d_{12}) + w_2 S(d_{22}) + w_3 S(d_{23}) &= S(d_{2p}) \\ w_1 S(d_{13}) + w_2 S(d_{23}) + w_3 S(d_{33}) &= S(d_{3p}) \end{aligned} \tag{4}$$

where $S(d_{ij})$ is the model variogram evaluated at a distance equal to the distance between points i and j. For example, $S(d_{1p})$ is the model variogram evaluated at a distance equal to the separation of points P_1 and P. Since it is necessary that the weights sum to unity, a fourth equation:

$$w_1 + w_2 + w_3 = 1.0 \tag{5}$$

is added. Since there are now four equations and three unknowns, a slack variable, λ, is added to the equation set. The final set of equations is as follows:

$$\begin{aligned} w_1 S(d_{11}) + w_2 S(d_{12}) + w_3 S(d_{13}) + \lambda &= S(d_{1p}) \\ w_1 S(d_{12}) + w_2 S(d_{22}) + w_3 S(d_{23}) + \lambda &= S(d_{2p}) \\ w_1 S(d_{13}) + w_2 S(d_{23}) + w_3 S(d_{33}) + \lambda &= S(d_{3p}) \\ w_1 + w_2 + w_3 + 0 &= 1.0 \end{aligned} \tag{6}$$

The equations are then solved for the weights w_1, w_2, and w_3. The f value of the interpolation point is then calculated as:

$$f_p = w_1 f_1 + w_2 f_2 + w_3 f_3 \tag{7}$$

By using the variogram in this fashion to compute the weights, the expected estimation error is minimized in a least squares sense. For this reason, kriging is sometimes said to produce the best linear unbiased estimate (BLUE). However, minimizing the expected error in a least squared sense is not always the most important criteria and in some cases, other interpolation schemes give more appropriate results (Philip & Watson, 1986).

An important feature of kriging is that the variogram can be used to calculate the expected error of estimation at each interpolation point since the estimation error is a function of the distance to surrounding scatter points. The estimation variance can be calculated as:

$$s_z^2 = w_1 S(d_{1p}) + w_2 S(d_{2p}) + w_3 S(d_{3p}) + \lambda \tag{8}$$

When interpolating to an object using the kriging method, an estimation variance data set is always produced along with the interpolated data set. As a result, a contour or iso-surface plot of estimation variance can be generated on the target mesh or grid.

Since kriging is a rather complex interpolation technique and includes numerous options, a complete description of kriging is beyond the scope of this report. The reader is strongly encouraged to consult the UNCERT User Guide (Wingle, *et al.*, 1995) and the GSLIB textbook (Deutsch and Journel, 1992) for more information. Other good references on kriging include Royle *et al.* (1981), Davis (1986), Lam (1983), Heine (1986), Olea (1974), Journel & Huijbregts (1978). Isaaks and Srivastiva's (1989) chapter on "Ordinary Kriging" is particularly helpful.

The preceding description is provided by Environmental Modeling Systems, Inc. (ems-i) through their website at http://www.ems-i.com. Figure 4 provides an example of the type of result generated by this modeling approach.

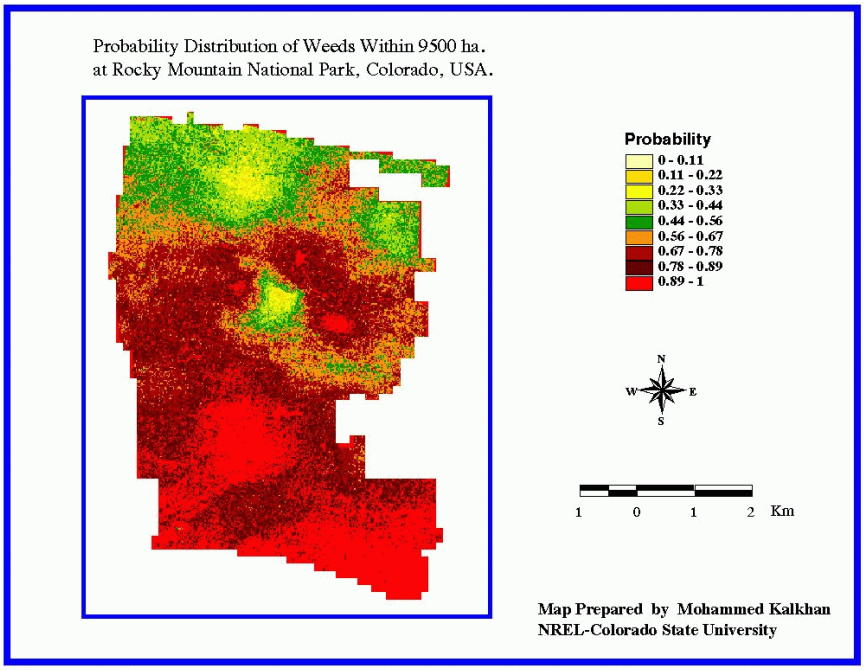


Figure 4. Predicted probability distribution of weeds within 9500 ha. at the Rocky Mountain National Park, Colorado, USA.

4 Software Description

The practical implementation of the modeling process, as developed and used by colleagues at USGS, consists of a pre-processing step, a modeling step, and a post-processing step.

4.1 Pre-Processing

Pre-processing activities merge ingested datasets to create a data product that can be analyzed in the subsequent modeling step. In the baseline scenario, the field data are merged with the Landsat and DEM information at the same UTM x,y coordinates. Resampling may be performed at this time if the input data are not at the same resolution, and the Landsat data may be processed to higher level products, e.g. tassel cap coefficients, principal components, atmospherically corrected reflectance values, etc. The merged data product is written to backing store in a common analysis format.

For the most part, this pre-processing step is a straightforward application of common techniques, and is thus not a major focus of our current work. A possible exception to be explored in the future is the pre-processing that might be needed to accommodate new data sets in the model, e.g., atmospherically corrected hyperspectral data which presents different types of computational challenges.

4.2 Modeling[1]

The primary modeling pipeline uses the merged, flat file resulting from the pre-processing step. The file is logically arranged with one row of data for each field survey point. The data sampled at each point are arranged in columns. The file contains a simple internal ASCII header that contains the number of rows and columns along with a one-word column descriptor. The columns include a subset of the following data: location, plant, soil, digital elevation model (DEM), and remote sensing information such as Landsat DN and derived quantities (e.g. tassel cap coefficients or NDVI values). A series of statistical analyses are then performed in S-plus as follows:

1. Read the input field data to create an object within S-plus.
2. Compute the distance matrix, which is the Euclidean distance between each sample point.
3. Perform a stepwise multiple regression with total plants as the dependent variable and the DEM, remote sensing, soils, etc. data as the independent variables.
4. Perform a weighted ordinary least squares (OLS) fit to the total plants for the independent variables that are found to be significant predictors.
5. Compute Moran's I coefficient to determine if there is spatial structure in the residuals of this OLS fit.
6. If there is no spatial structure then skip to step 10.
7. Compute the variogram of the residuals to determine the spatial structure.
8. Determine whether a gaussian, exponential, and spherical model best fits the variogram.

[1]The processing flow described here was developed by USGS. We have developed a wrapper that automates these steps in order to execute baseline runs. See Appendix C for further details.

9. Perform kriging to estimate the residual surface across the entire study area. The kriging can be performed either in S-plus or using a FORTRAN program.

   (a) If S-plus kriging is performed:

      i. The kriged residual surface is created directly as an S-plus object. Error estimates are also calculated.

      ii. The kriged residuals and estimated errors are then written to separate ASCII files. These files contain headers listing the number of rows and columns in the kriged surface along with georeferencing information.

   (b) If FORTRAN kriging is performed:

      i. The residuals at each field sample point are written to an intermediate ASCII file, along with a header containing parameters that control the kriging. These parameters include the number of rows and columns to be kriged, the spatial resolution and georeferencing information for the kriged surface, the number of field sample points, the number of nearest neighbors to include in the estimation, the parameters describing the variogram model (e.g. range, nugget, sill, gaussian), and finally a flag indicating whether to calculate the error estimates for the kriged residuals.

      ii. The compiled and linked FORTRAN executable is invoked. It reads the above file of residuals, performs the kriging, and writes the kriged residuals to an output ASCII file. The format of this file is identical to that created if S-plus kriging is performed. The error estimates for the kriged residuals are written to a separate file using the same format.

   (c) The ASCII files with the results of the kriging are converted to a binary raster format using a simple filter. Separate header files are created so that the kriging results can be easily viewed the using version 3.5 of ENVI, the Environment for Visualizing Images from Research Systems, Inc. (see http://www.rsinc.com).

10. Apply the results of the OLS fit to the maps of the significant independent variables to create an estimate of the spatial distribution of the total plants.

11. If the residuals were kriged in step 9, then add the kriged residual surface to the estimate of the total plants.

4.3 Post-Processing

The post-processing step applies the results of the above modeling activities to generate products such as the map shown in Figure 5. In the canonical case, we:

12. Create a JPEG rendering of the total plants map. Steps 10 - 12 are performed using version 5.5 of IDL, the Interactive Data Language from Research Systems, Inc.

13. Create separate header files so that the total plants and error estimates can be easily read using ENVI.

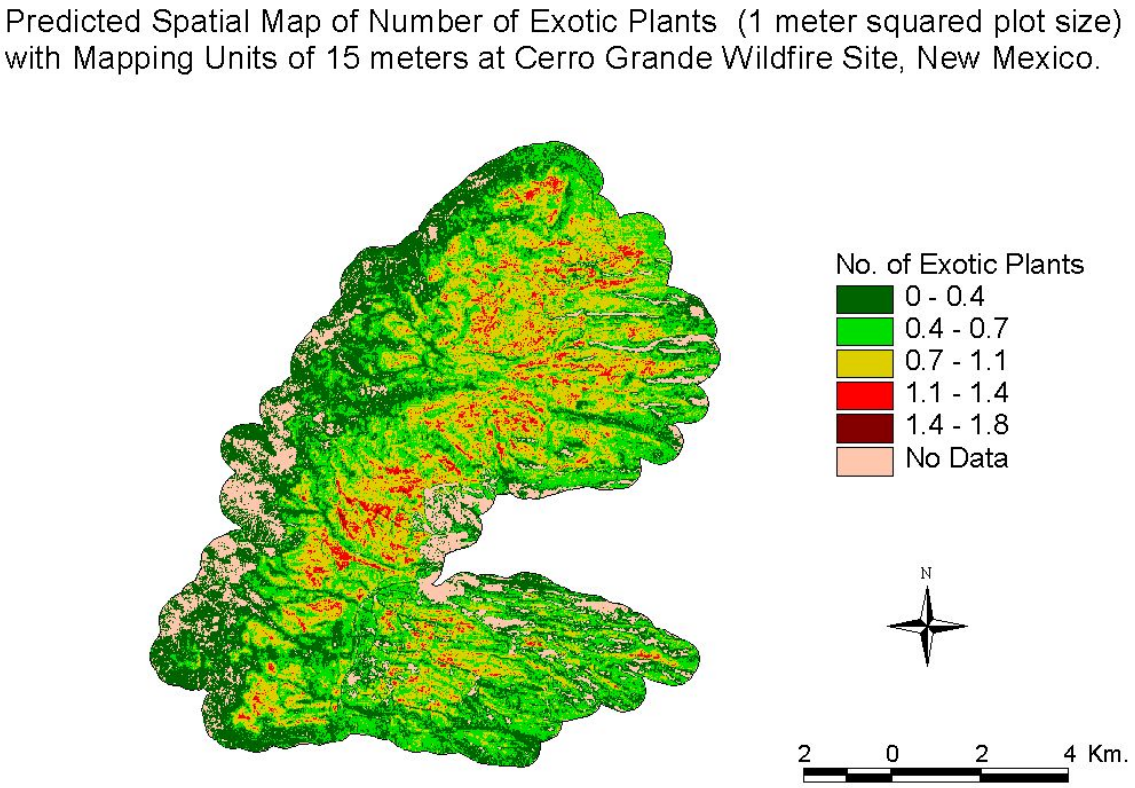


Figure 5. Predicted Spatial Map of Exotic Plants at Cerro Grande Wildfire Site

5 Baseline Scenario

In this project, we are working with three "canonical" study sites: the Cerro Grande Fire Site in Los Alamos, NM (CGFS), Rocky Mountain National Park, CO (RMNP), and Grand Staircase Escalante National Monument, UT (GSENM). The three sites provide contrasting ecological settings and analysis challenges and vary in the types and scales of data used, areas covered, and maturity of the investigation.

In many respects, the most comprehensive modeling efforts to date have involved the Cerro Grande site which covers approximately 50,000 acres. Cerro Grande modeling activities integrate a range of environmental attributes and data sets including, as described below, data from over 1000 field sample plots. The Rocky Mountain study site, at nearly 25,000 acres, is a smaller study area but includes data from over 1000 field sample sites. RMNP has been studied the longest of the three sites. The newest study area is Grand Staircase Escalante National Monument, which at 1.9 million acres is by far the largest investigation to be undertaken with these modeling techniques. The gathering of field data started three years ago and continues with over 350 plots being studied to date.

Each study site will be used to examine various aspects of the modeling approach. In terms of number of sample plots and output coverage, Cerro Grande represents a typical dataset and provides the best opportunity to analyze baseline performance characteristics of the modeling system. RMNP is an example of a dataset with a relatively large number of sample plots and a fairly small output surface. GSENM, on the other hand, is an example of a where sample plots are relatively few, but coverage area is large. Field work on the GSENM study site will not be completed until the end of next summer. We have thus chosen CGFS and RMNP as the two canonical examples to be used in establishing baseline performance characteristics of the model. We provide a comprehensive analysis for CGFS and use CGFS as the basis for defining our community goals for code improvement. We then provide a summary analysis for RMNP and use RMNP as the basis for defining some advance applications goals.

5.1 CGFS and RMNP Study Areas

Investigating spatial relationships among fuels, wildfire severity, and post-fire invasion by exotic plant species through linkage of multi-phase sampling design and multi-scale nested sampling field plots, pre- and post fire, has been accomplished on the CGFS using the *PlantDiversity* model. The technique provides useful information and tools for describing ecological and environmental characteristics including landscape-scale fire regimes, invasive plants, and hot spots of diversity (native and non-native plants) for the site. Data from the Rocky Mountain National Park (RMNP) study site likewise have been used to predict the distribution, presence, and patterns of native and exotic plants with a focus on providing land managers with better techniques to assess native biodiversity and the potential for exotic invasions.

The Cerro Grande Fire Site is located near Los Alamos, New Mexico with elevation range from 1932 to 3200 meters. The Cerro Grande fire began as a prescribed fuel treatment by Bandelier National Monument, Los Alamos, NM on 4 May 2000. The fire escaped control and was declared a wildfire on 5 May 2000. The fire was contained on 24 May after burning about 19,300 ha of lands managed by seven different agencies, including the town of Los Alamos, NM. However, 60% of the fire area burned 10-11 May, 2000, and 60% of the fire was on the Espaola District of the Santa Fe National Forest (Burned Area Emergency Rehabilitation [BAER] Team 2000). Initial remotely sensed estimates of burn severity were classified as high (35%), moderate (9%), and low (56%).

To predict the distribution, presence, and patterns of native and exotic species in, we used data points (based on Modified-Whittaker nested plots of 1000 m2) to represent different variables that were extracted from Landsat TM data (eight bands, six vegetation indices, and six bands of tasseled cap

transformation indices), topographic data (elevation, slope, and absolute aspect), and vegetation characteristics. A total of 79 data points were used with CGFS and a total of 1180 points for RMNP. Spatial statistics were used to integrate these data to model large- and small-scale variability. In the canonical case, we use trend surface models that describe the large-scale spatial variability using stepwise multiple regressions based on the Ordinary Least Squares (OLS) method. Models with small variance were selected. In addition, the residuals from the trend surface model based on the OLS estimates were modeled using ordinary kriging for modeling small-scale variability based on a Gaussian semi-variogram. The final surfaces were obtained by combining two models (the trend surface based on the OLS and the kriging surface of residuals). All models were selected based on lowest values of standard errors, AICC statistics, and high R2 . For large-scale spatial variability models using the OLS procedure, R2 values ranged from 10.04% to 58.6% in the CGFS data and all variables were significant at $\alpha < 0.05$ level. When adding the kriging model with the OLS model, R2 values ranged from 60% to 84% for CGFS. Similar results have been obtained with the RMNP data.

5.2 Software and Hardware Environment

The baseline processing system uses a combination of COTS and public domain software to generate maps of estimated biodiversity or ecosystem parameters. The major software components include the following:

- S-plus version 6.0.1 for Linux has been used for the baseline test cases. S-plus is a commercially available statistical package commonly used in many scientific communities. The functionality of S-plus has been enhanced by a large collection of spatial statistical functions developed and maintained by Drs. Robin Reich and Richard Davis of the Colorado State University (see http://www.cnr.colostate.edu/ robin/).
- Version 3.5 of ENVI, the Environment for Visualizing Images from Research Systems, Inc. (see http://www.rsinc.com). ENVI is a commercially available image analysis application in common use in the remote sensing community.
- Version 5.5 of IDL, the Interactive Data Language from Research Systems, Inc. IDL is a commercially available image analysis application in common use in the remote sensing community.

The baseline processing hardware is a single-processor AMD machine running Redhat Linux 7 (kernel 2.4.9-31) at 1.2 GHz with 1.5 GB of RAM and 60 GB of disk space.

5.3 Performance Characteristics

Three factors influence the performance of the *PlantDiversity* model: the size of the output surface area over which kriging occurs (area), the total number of number of sample points in the data set (pts), and the number of "nearest neighbor" (nn) sample points from the total data set actually used to compute a kriged value for any given point in the output area. When we first began work with colleagues at USGS, a scalar, single-processor run of this model using S-plus took approximately two weeks. The major computational bottleneck in the model is the kriging routine. Solving for the weights in the equations that form the ordinary kriging system (Eq. 6) uses LU decomposition with backsubstitution to do matrix inversions. The overall computational complexity of ordinary kriging is thus O(n^3), and the time required to compute a result is strongly influenced by the number of sampled data points used to estimate the residual surface across the entire study area.

In order to achieve the two-week result described above, initial USGS model runs limited kriging (and thus the size of the computed covariance matrices) to only 18 out of the total 79 sample points for Cerro Grande. The kriging process iterates over the rows and columns of the output surface. For each (i,j) point of the output surface area, 18 nearest neighbor sample points were found and ultimately transformed into the appropriate weighted average to estimate the kriged value at point (i,j). This sub-sampling of 18 nearest neighbors is significant because it represents an accommodation that may be appropriate and exploitable in some circumstances while other types of applications may require the use of significantly more sampled points or the entire set of sampled points. The implications of these options will be explained in more detail below.

The output surface representing the entire CGFS area consists of 652 x 715 points, and the canonical CGFS dataset consists of 79 data points. The output surface representing the entire RMNP area consists of 1186 x 1041 points and the canonical RMNP dataset consists of 1180 points. Appendix B and Figures 6 and 7 show the results of baseline runs of the CGFS case and confirm that processing time scales linearly with both the size of the output surface (area) and the number of total data points (pts). In contrast, processing time increases order n^3 with respect to the number of nearest neighbors (nn) being used in the kriging routine (Figure 8). In order to do our baseline run, a FORTRAN kriging routine was developed that ran approximately two orders of magnitude faster than the original S-plus routine (see Appendix D). While the overall project of growing *PlantDiversity* into a comprehensive *Invasive Species Forecasting System* involves many elements, kriging will continue to be the focus of our efforts to improve model performance.

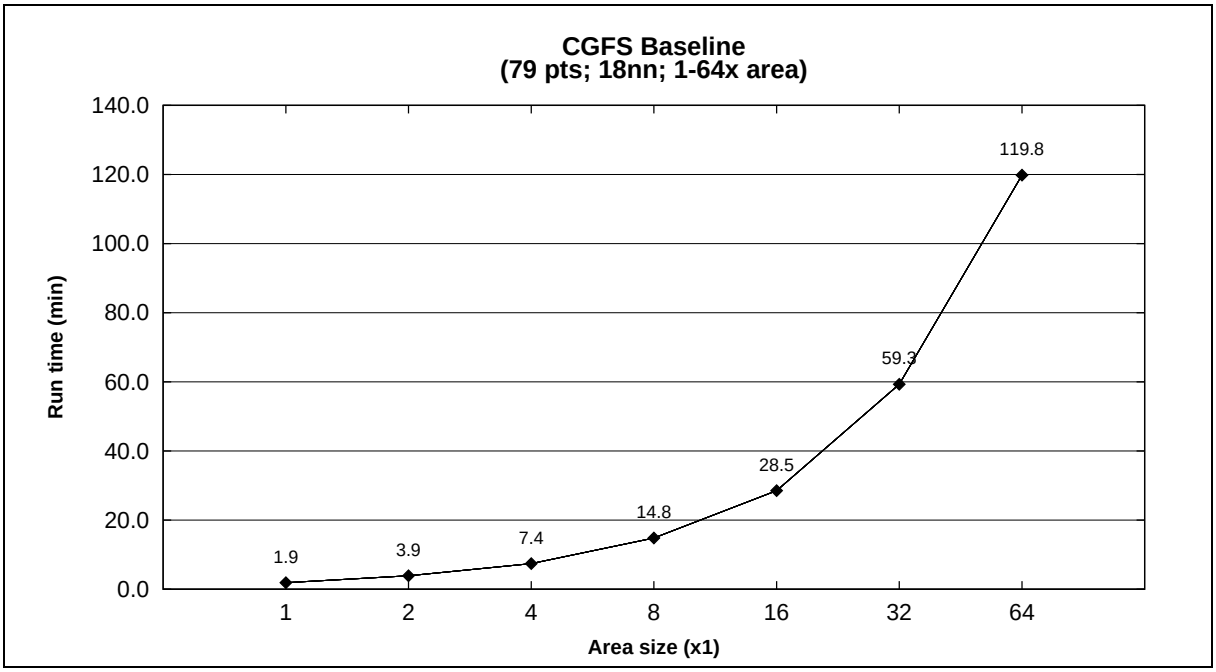


Figure 6. Graph of Runtime vs. Area size for CGFS Baseline

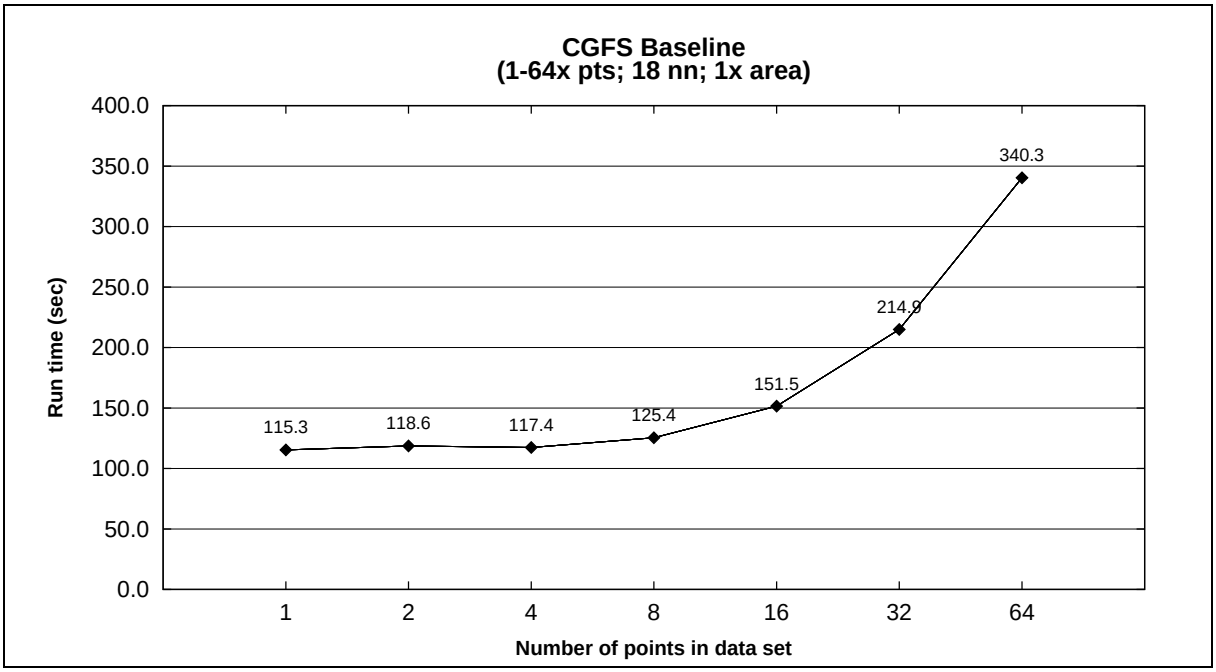


Figure 7. Graph of Runtime vs. Data Set Size for CGFS Baseline

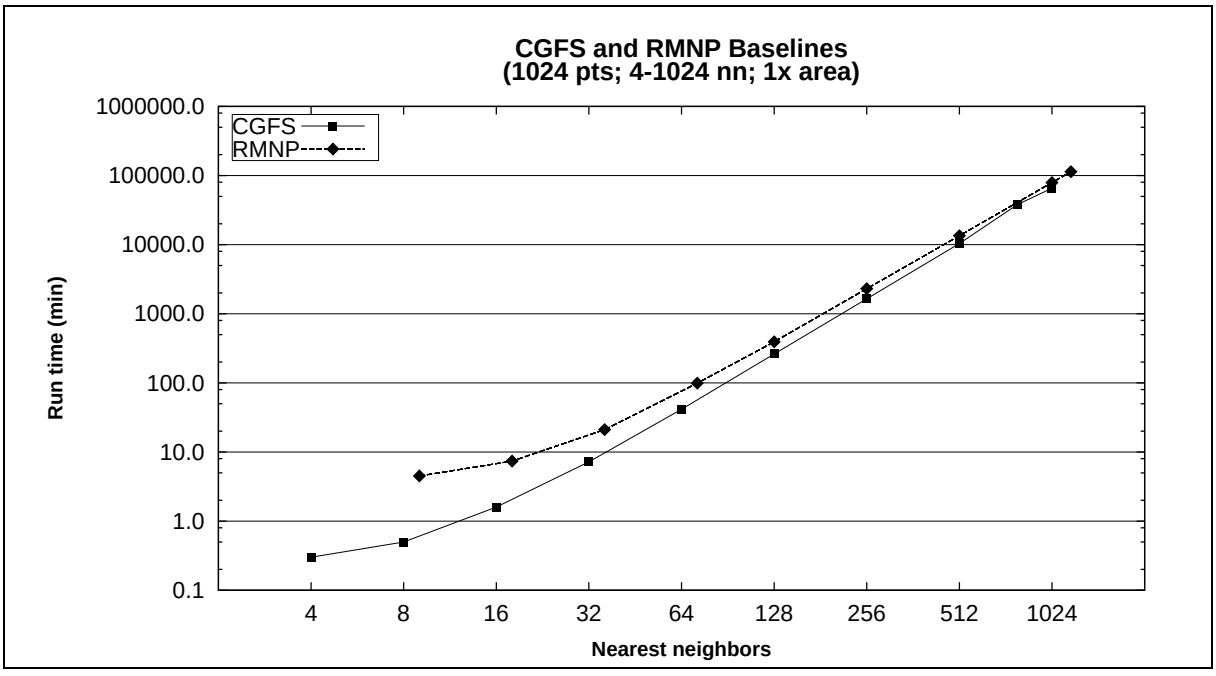


Figure 8. Graph of Runtime vs. Number of Nearest neighbors used for CGFS and RMNP Baselines

6 Performance Improvement Plan

6.1 Parallelization Strategy

The overall goal for code improvement is to reduce processing times and increase the amount of data handled by the model. As described below, increasing the amount of data handled by the model translates into either increasing spatiotemporal resolution or increasing coverage (Table 6.1). We first wish to accomplish quantitative improvements in the underlying model that have been agreed upon by the user community as minimal advances needed to improve core capabilities. These goals were driven by the knowledge that in Year 2 of the project we will build a 32-node cluster in the USGS facility. We refer to these as "Community Improvement Goals." The ESTO/CT program, however, provides access to even greater computational capabilities that can be used to apply this modeling approach to some important and challenging problems that have heretofore been unapproachable. We would therefore like to use CT's clusters to attain more challenging performance improvement goals at the same time we are accommodating basic needs. We refer to these complimentary challenges as "Advanced Improvement Goals." Each will be described below.

Table 2. Current performance characteristics and improvement goals.

BASELINE SCENARIO		**Sec**	**Min**	**Hrs**	**Days**	
CGFS_base_079_pts_79_nn_01x_area (S-Plus) (Version 0.0)		-	-	-	-	
CGFS_base_079_pts_18_nn_01x_area (S-Plus) (Version 0.0) (USGS Actual)		1209600.0	20160.0	336.0	14.0	
CGFS_base_079_pts_18_nn_01x_area (S-Plus) (Version 0.0) (NASA Estimate)		1608426.0	26807.1	446.8	18.6	
CGFS_base_079_pts_18_nn_01x_area (FORTRAN) (Version 0.1)		114.5	1.9	0.0	0.0	
CGFS_base_079_pts_79_nn_01x_area (FORTRAN) (Version 0.1)		4702.6	78.4	1.3	0.1	A
RMNP_base_1180_pts_18_nn_01x_area (FORTRAN) (Version 0.1)		443.0	7.4	0.1	0.0	
RMNP_base_1180_pts_1180_nn_01x_area (FORTRAN) (Version 0.1) (est.)		6812384.0	113539.7	1892.3	78.8	B
COMMUNITY IMPROVEMENT (CI) GOALS	**x baseline**	**Sec**	**Min**	**Hrs**	**Days**	
CGFS_base_079_pts_79_nn_01x_area (Version 1.0 - F)	25.0	188.1	3.1	0.1	0.0	C
CGFS_base_790_pts_79_nn_01x_area (Version 2.0 - G)	25.0	188.1	3.1	0.1	0.0	D
CGFS_base_790_pts_79_nn_10x_area (Version 2.0 - G)	2.5	1881.0	31.4	0.5	0.0	E
ADVANCED IMPROVEMENT (AI) GOALS	**x baseline**	**Sec**	**Min**	**Hrs**	**Days**	
CGFS_base_079_pts_79_nn_01x_area (Version 1.0 - F)	200	23.5	0.4	0.0	0.0	F
RMNP_base_1180_pts_1180_nn_01x_area (Version 2.0 - G)	1000.0	6812.4	113.5	1.9	0.1	G
RMNP_base_11800_pts_1180_nn_01x_area (Version 2.0 - G)	1000.0	6812.4	113.5	1.9	0.1	H
RMNP_base_11800_pts_1180_nn_100x_area (Version 2.0 - G)	10.0	681238.4	11354.0	189.2	7.9	I

A Proposed CGFS canonical baseline using FORTRAN kriging routine.
B Proposed RMNP canonical baseline using FORTRAN kriging routine.

C Milestone F CI Goal - speed up - 75% efficiency, 32 node cluster = 25x speed up
D Milestone G CI Goal - increased resolution - "sliding window" adaptive selection of 10% of 10x nn from 1x area
E Milestone G CI Goal - increased coverage - "sliding window" adaptive selection of 10% of 10x nn from 10x area

F Milestone F AI Goal - speed up - 75% efficiency, 256+ node cluster = 200x speed up
G Milestone G AI Goal - speed up - 75% efficiency, 1024+ node cluster = 1000x speed up
H Milestone G AI Goal - increased resolution - "sliding window" adaptive selection of 10% of 100x nn from 1x area
I Milestone G AI Goal - increased coverage - "sliding window" adaptive selection of 10% of 100x nn from 10x area

6.2 Milestone F — First Code Improvement (Parallelization)

Improve implementation of PlantDiversity to deliver canonical products from Milestone E mX faster than the baseline implementation. Provide code scaling curves. Deliver updates to Requirements and Design Documents. Deliver initial version of Test Plan / Procedures Document. Documented source code made publicly available via the Web. Complete Optional Milestone.

For Milestone F, we will focus on the community goal of increasing speed by at least a factor of 25. This corresponds to "$m = 25$" in our Milestone F Goal Statement. As a practical matter, this would mean that it would be possible to reduce the time to process the CGFS dataset from over one hour to less than five minutes.[2] For the advanced application goal, an important component of the PlantDiversity model is its consideration of the source of error in the application of the model. We would like to use Monte Carlo simulation methods to examine the effects of error propagation. We are also developing scenarios where the system can be used for interactive "what-if" explorations of data subsets. We believe that this level of interactivity could be attained if it were possible to maintain the proposed 75% scaling efficiency to 256-node or larger clusters. Doing so would translate into at least a 200x improvement in the baseline codes and allow a the 79-point CGFS dataset to be kriged in less than half a minute. This corresponds to "$m = 200$" in our Milestone F Goal Statement for this advanced application.

Our approach will be to develop parallel, Message Passing Interface (MPI) versions of the *PlantDiversity* model. In *PlantDiversity*, the kriged estimate for each pixel of the output surface does not depend on neighboring pixels. The kriging process is thus spatially independent, and we expect to achieve good scaling performance by domain decomposition. A parallel implementation of kriging would have three steps. The first step would broadcast input data from a "control node" to all other nodes of a cluster. The second step would have each node compute its piece of the kriged surface. For example, node 1 could process rows 1 - 16, node 2 could process rows 17 - 32, etc. The final step would assemble the entire kriged surface on the control node. The parallel second step is maximally efficient; the first and last serial steps are the overhead costs of a parallel implementation.

Achieving a 25x speed up will require approximately 75% scaling efficiency on a 32 node cluster, or its equivalent on a larger cluster[3]. This efficiency requires that we krig the CGFS residuals in about 3.0 minutes. A completely efficient implementation would give a run time of about 2.5 minutes. We must therefore incur no more than 30 seconds of parallel overhead to meet our goal. We feel this is attainable since even a serial broadcast of the input data using 'scp', followed by a serial gather of the individual pieces of the kriged surface, can be accomplished in this time on a 10base-T network. A cluster built with 2Gbps Myrinet and using much lower overhead communications calls should be able to perform substantially better. Further, it should be possible to hide nearly all the step three communications overhead by overlapping the computation of the current row with the sending of results from the previous row to the control node. In this approach each processor would calculate the first row, asynchronously send a message to the control node with the results of the first row, then immediately continue processing the second row. Synchronization would only be required at the end of the second row to check that the results of the first row had been received successfully. This approach should be efficient since only a small number (≈ 31) of fairly large messages (≈ 2.6 KB) need to be sent each time a row has been calculated (≈ 5.7 seconds).

[2] Note that while the original S-plus version of the model took 14 days to run, we are using the single-processor FORTRAN version of the kriging routine as our canonical baseline.

[3] As part of this project, we will build and deploy a cluster of at least 32 nodes at USGS specifically to support these modeling activities.

6.3 Milestone G — Second Code Improvement (Adaptive Kriging)

Improve implementation of PlantDiversity to accommodate 10X more input data over Milestones E and F at nX the time required in the baseline implementation. (Depending on the science problem, this enhanced capability may be used to increase spatial resolution, temporal resolution, or coverage.) Provide code scaling curves. Deliver updates to Requirements, Design, and Test Documents. Deliver initial User's Guide. Documented source code made publicly available via the Web.

There are two ways that the model can be construed to accommodate more data: the model can either handle more data points in its kriging routine or it can krig data over a larger area. For both cases, we propose to combine the improvements achieved through parallelization with an adaptive approach to sub-sampling datasets.

In natural systems, spatial processes have a finite range of influence. As the number of sample data points grows (say from 79, in the case of CGFS, to 10x = 790), we do not necessarily wish to scale the kriging procedure to the full set of sampled data. (In fact, if the entire data set is used, computing the covariance matrix need only be done once, vastly reducing the computational complexity of the kriging task.) Ideally, we would rather scale on the basis of the number of spatially-relevant nearest neighbors in the region we wish to estimate. In many cases, this will involve only a fraction of a larger dataset. In the original applications of the model, USGS arbitrarily kriged using only 18 of 79 sample points based on a simple notion of nearest neighbor Euclidian distance of sample points from points of the output surface. We propose to refine these techniques by using a "sliding window" that at each point of the output surface adaptively selects a subset of sample points for kriging based on statistics or on an understood spatial influence on the physics or biology of the dependant variables being examined. We refer to this approach as "adaptive kriging."

Intelligent minimization of the number of nearest neighbors can significantly improve the model's ability to handle larger datasets. Perhaps more important, it provides a context for exploring mechanistic aspects of the modeling problem that are elided by the overarching statistical approach. Assuming that adaptive kriging uses 10% of the total number of sampled points, a 10x increase in dataset size essentially returns us to the baseline condition where we would expect a 25x speed-up through the Milestone F parallelization. Since the problem scales linearly with area, performing adaptive kriging over a ten-fold larger output area would result in an expected speed-up of approximately 2.5x. These cases correspond respectively to "$n = 25$" and "$n = 2.5$" in our Milestone G Goal Statement for the 32-node processor community improvement goal.

The Milestone G advanced improvement goal will position the project to accommodate a new suite of data sources that will become available soon. County-level plot data are currently being gathered from a range of sources for the entire state of Colorado (66.7 million acres), and decadal-scale time series data on the spread of several invasive species in the Southwestern US (500 million acres) are being assembled. These will comprise thousands of field measurements. Within the next year, we believe it will be possible to use 1024-node and larger clusters running this modeling system to take on state- and regional-scale problems such as these. Here, the goal would be to accommodate 100x more points over as much as 100x more output surface than in the canonical examples. Our empirical measure of time complexity for baseline (Figure 8) suggests that a single-processor run would take approximately one month to krig 1000 points in a large dataset such as RMNP. Reducing the processing time to one hour or less would dramatically improve capabilities and could be accomplished if 75% scaling efficiency could be preserved over 1024 processors or more. We would therefore like to make this our second and most challenging advanced application goal thus corresponding to "$n = 1000$" or "$n = 10$" for 100x data in

our Milestone G Goal Statement.

It is important to note that adaptive kriging using more data over a fixed-size area essentially increases model spatial resolution; adaptive kriging using a fixed-size dataset over a larger output area increases model spatial coverage; in both cases, repeated runs of adaptive kriging using time-series data increases the model's temporal resolution. All three classes of improvement have been identified by the research community as needed enhancements to the science and technology underlying biotic prediction.

A summary of the proposed improvement goals for our project are shown in table 3.

Table 3. Summary of Proposed Improvement Goals

Milestone	**Community Goal (32-node cluster)**	**Advanced Applications Goal (256[a]- & 1024[b]-node clusters)**
F — Parallelization	25x speed $(m = 25)$	200x speed $(m = 200)$[a]
G — Adaptive Kriging	10x data $(n = 25; n = 2.5)$	1000x speed $(m = 1000)$[b] 100x data ($n = 1000$; $n = 10$)

7 Baseline Software / System Delivery

The baseline system along with complete documentation are available on the project's BPDEV computer ("frio.gsfc.nasa.gov"). ESTO/CT Milestone E deliverables for this project are available at http://ltpwww.gsfc.nasa.gov/BP/deliverables.html. Users may log on to the system to run the baseline program (please contact John Schnase at 6-4351 for userid and password). In addition, a tarfile is available from both the website and the ISFS home directory that can be used to build the baseline environment on a different machine.

8 References

Agee, J. K. and D. R. Johnson. 1988. Ecosystem management for parks and wilderness. University of Washington Press, Seattle.

Association for Computing Machinery [ACM]. 2000. *Communications of the ACM*, Special Issue on Component-Based Enterprise Frameworks, Vol. 43, No. 10.

Bonham, C. D., R. M. Reich, and K. K. Leader. 1995. A spatial cross-correlation of *Bouteloua gracilis* with site factors. *Grasslands Science*, Vol. 41, pp. 196-201.

Chong, G. W., R. M. Reich, M. A. Kalkhan, and T. J. Stohlgren. 2000. New approaches for sampling and modeling native and exotic plant species richness. *Western North American Naturalist* (in review).

Czapleski, R. L. and R. M. Reich. 1993. *Expected value and variance of Moran's bivariate spatial autocorrelation statistic under permutation*, Research Paper RM-309. U. S. Department of Agriculture, Rocky Mountain Experimental Range Station, Fort Collins, CO.

Glass, G. E. 2000. Geographic information systems. In: K. Nelson, C. Masters, and N. Graham, *Infectious Disease Epidemiology*, Aspen Publishing, pp. 231-253.

Gown, S. N., R. H. Waring, D. G. Dye, and J. Yang. 1994. Ecological remote sensing at OTTER: Satellite Macroscale Observation. *Ecological Applications*, Vol. 4, pp. 322-343.

Isaaks, E. H. and R. M. Srivastava. 1989. *An Introduction to Applied Geostatistics*. Oxford University Press, New York.

Kalkhan, M. A. and T. J. Stohlgren. 2000. Using multi-scale sampling and spatial cross-correlation to investigate patterns of plant species richness. *Environmental Monitoring and Assessment*, Vol. 64, No. 3, pp. 591-605.

Kalkhan, M. A., T. J. Stohlgren, and M. B. Coughenour. 1995. An investigation of biodiversity and landscape-scale gap patterns using double sampling: a GIS approach. *In: Ninth International Symposium on Geographic Information Systems for Natural Resources, Environment, and Land Information Management*, pp. 708-712.

Kalkhan, M. A., R. M. Reich, and T. J. Stohlgren. 1998. Assessing the accuracy of Landsat Thematic Mapper classification using double sampling. *International Journal of Remote Sensing*, Vol. 19, pp. 2049-2060.

Kalkhan, M. A., G. W. Chong, R. M. Reich, and T. J. Stohlgren. 2000a. Landscape-scale assessment of plant diversity under mountain terrain: integration of remotely sensed data, GIS and spatial statistics. *In*: ASPRS Annual Convention and Exposition, ASPRS Technical Papers, May 22-26, 2000, Washington, DC.

Kalkhan, M. A., T. J. Stohlgren, G. W. Chong, L. D. Schell, and R. M. Reich. 2000b. A predictive spatial model of plant diversity: integration of remotely sensed data, GIS, and spatial statistics. *In*: Eight Biennial Remote Sensing Application Conference (RS 2000), April 10-14,2000, Albuquerque, New Mexico, (In press).

Kalkhan, M. A, T. J. Stohlgren, and G. W. Chong 2000c. Landscape-scale Assessment of Plant Diversity: Integration of Spatial Information and Spatial Statistics. *In*: 85rd Annual Meeting Ecological Society of America (Contributed Papers: Landscape Ecology), August 6-10, 2000, Snowbird, Utah, (In press).

Kallas, M. 1997. *Hazard Rating of Armillaria Root Rot on the Black Hills National Forest*. M. S. Thesis, Department of Forest Sciences, Colorado State University, Fort Collins, CO.

LaRoe, E.T. 1993. Implementation of an ecosystem approach to endangered species conservation. *Endangered Species Update*, Vol. 10, pp. 3-6.

McNaughton, S. J. 1993. Biodiversity and function of grazing ecosystems. *In*: Schulze, E-D. and H. A. Mooney, *Biodiversity and ecosystem function*, Ecol. Studies 99. Springer-Verlag, Berlin, pp. 361-408.

Metzger, K. 1997. *Modeling small-scale spatial variability in stand structure using remote sensing and field data*. M. S. Thesis, Department of Forest Sciences, Colorado State University, Fort Collins, CO.

Morrill, W. L., 1998. Production and flight of alate red imported fire ants. *Environmental Entomology*, Vol. 3, pp. 265-271.

National Research Council [NRC]. 2000. *Grand Challenges in Environmental Sciences*. Report of the National Research Council Committee on Grand Challenges in Environmental Sciences, National Academy Press, Washington, DC. 88 pp.

Noss, R. 1983. A regional landscape approach to maintain diversity. *BioScience*, Vol. 33, pp. 700-706.

Office of Science and Technology Policy Committee on Environment and Natural Resources [OSTP/CENR]. 2000. *Ecological Forecasting and Integrating Science for Ecosystem Challenges (in press.)*

Paulk, M., B. Curtis, M. B. Chrissis, and C. Weber. 1993. *Capability Maturity Model for Software, Version 1.1* (CMU/SEI-93-TR-024, ADA 263403). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University.

President's Committee of Advisors on Science and Technology [PCAST]. 1998. *Teaming with Life: Investing in Science to Understand and Use America's Living Capital*. Report of the PCAST Panel on Biodiversity and Ecosystems.

Reich, R. M., C. Aguirre-Bravo, M. A. Kalkhan, and V. A. Bravo. 1999. Spatially based forest inventory and monitoring system for Ejido El Largo, Chihuahua, Mexico. *In: North America Science Symposium Toward a Unified Framework for Inventorying and Monitoring Forest Ecosystem Resources* (Contributed Papers: Quantitative Spatial Analysis Applications), November 1-6, 1998, Guadalajara, Jalisco, Mexico. pp. 31-41.

Reich, R. M., R. L. Czaplewski, and W. A. Bechtold. 1994. Spatial cross-correlation in growth of undisturbed natural shortleaf pine stands in northern Georgia. *J. Environmental and Ecological Statistics*, Vol. 1, pp. 201-217.

Robertson, G. P. 1987. Geostatistics in ecology: interpolating with known variance. Ecology, Vol. 63, pp. 744-748.

Schnase, J. L. 2000. Research directions in biodiversity informatics. *In: Proceedings of the Very Large Databases Conference* (VLDB 2000), Cairo, Egypt, August 9-15, pp. 217-220.

Schnase, J. L., D. K. Kama, K. L. Tomlinson, J. A. Snchez, E. L. Cunnius, and N. R. Morin. 1997. The Flora of North America digital library: A case study in biodiversity database publishing. *Journal of CNetwork and Computer Applications*, Vol. 20, pp. 87-103.

Schnase, J. L., M. A. Lane, G. C. Bowker, S. L. Star, A. Silberschatz. 2000. Building the next-generation biological-information infrastructure. *In*: P.H. Raven, *Nature and Human Society: The Quest for a Sustainable World*, National Research Council, National Academy Press, Washington, DC, pp. 291-300.

Stohlgren, T. J., G. W. Chong, M. A. Kalkhan, and L. D. Schell. 1997a. Rapid assessment of plant diversity patterns: a methodology for landscapes. *Ecological Monitoring and Assessment*, Vol. 48, pp. 25-43.

Stohlgren, T. J., M. B. Coughenour, G. W. Chong, D. Binkley, M. Kalkhan, L. D. Schell, D. Buckley, and J. Berry. 1997b. Landscape analysis of plant diversity. *Landscape Ecology*, Vol. 12, pp. 155-170.

Stohlgren, T. J., G. W. Chong, M. A. Kalkhan, and L. D. Schell. 1997c. Multi-scale sampling of Plant Diversity: Effects of the minimum mapping unit. *Ecological Applications*, Vol. 7, pp. 1064-1074.

Stohlgren, T. J., D. Binkley, G. W. Chong, M. A. Kalkhan, L. D. Schell, K. A. Bull, Y. Otsuki, G. Newman, M. Bashkin, and Y. Son. 1999a. Exotic plant species invade hot spots of native plant diversity. *Ecological Monographs*, Vol. 69, pp. 25-46.

Stohlgren, T. J., L. D. Schell, and B. Vanden Heuvel. 1999b. Effects of grazing and soil characteristics on native and exotic plant species in Rocky Mountain grasslands. *Ecological Applications*, (in press).

Vinson, S. B. 1997. Invasion of the Red Imported Fire Ant. *American Entomologist*, Vol. 43, No. 1, pp. 265-271.

A Glossary

BP Biotic Prediction project
CGFS Cerro Grande Fire Site
CT Computational Technologies project
CONOP Concept of Operations
COTS Commercial Off The Shelf
CSU Colorado State University
ESTO Earth Science Technology Office
GSENM Grade Staircase Escalante National Monument
GSFC Goddard Space Flight Center
GUI Graphical User Interface
ISFS Invasive Species Forecasting System
NREL Natural Resources Ecology Laboratory
NDVI Normalized Differential Vegetation Index
RMNP Rocky Mountain National Park
SEP Software Engineering / Development Plan
URL Uniform Resource Locator

B CGFS and RMNP Detailed Performance Characteristics

		Sec	Min	Hrs	Days	
BASELINE SCENARIO		-	-	-	-	
CGFS_base_079_pts_79_nn_01x_area (S-Plus) (Version 0.0)		1209600.0	20160.0	336.0	14.0	
CGFS_base_079_pts_18_nn_01x_area (S-Plus) (Version 0.0) (USGS Actual)		1608426.0	26807.1	446.8	18.6	
CGFS_base_079_pts_18_nn_01x_area (S-Plus) (Version 0.0) (NASA Estimate)		114.5	1.9	0.0	0.0	
CGFS_base_079_pts_18_nn_01x_area (FORTRAN) (Version 0.1)		4702.6	78.4	1.3	0.1	* CGFS Canonical baseline
CGFS_base_079_pts_79_nn_01x_area (FORTRAN) (Version 0.1)		443.0	7.4	0.1	0.0	
RMNP_base_1180_pts_18_nn_01x_area (FORTRAN) (Version 0.1)		6812384.0	113539.7	1892.3	78.8	* RMNP Canonical baseline
RMNP_base_1180_pts_1180_nn_01x_area (FORTRAN) (Version 0.1) (est.)						
COMMUNITY IMPROVEMENT GOALS	**x baseline**	**Sec**	**Min**	**Hrs**	**Days**	
CGFS_base_079_pts_79_nn_01x_area (Version 1.0 - F) (inc. speed)	25.0	188.1	3.1	0.1	0.0	("75% efficiency, 32 node cluster = 25x speed up)
CGFS_base_790_pts_79_nn_01x_area (Version 2.0 - G) (inc. resolution)	25.0	188.1	3.1	0.1	0.0	("sliding window" to adaptively select 10% of 10x number of points from 1x area)
CGFS_base_790_pts_79_nn_10x_area (Version 2.0 - G) (inc. spatiotemporal coverage)	2.5	1881.0	31.4	0.5	0.0	("sliding window" to adaptively select 10% of 10x number of points from 10x area)
ADVANCED IMPROVEMENT GOALS	**x baseline**	**Sec**	**Min**	**Hrs**	**Days**	
CGFS_base_079_pts_79_nn_01x_area (Version 1.0 - F)	200.0	23.5	0.4	0.0	0.0	("75% efficiency, 256-node or larger cluster = 200x speed up)
RMNP_base_1180_pts_1180_nn_01x_area (Version 2.0 - G) (inc. speed)	1000.0	6812.4	113.5	1.9	0.1	("75% efficiency, 1024-node or larger cluster = 1000x speed up)
RMNP_base_11800_pts_1180_nn_01x_area (Version 2.0 - G) (inc. resolution)	1000.0	6812.4	113.5	1.9	0.1	("sliding window" to adaptively select 10% of 100x number of points from 1x area)
RMNP_base_11800_pts_1180_nn_100x_area (Version 2.0 - G) (inc. spatiotemorral coverage)	10.0	68238.4	11354.0	189.2	7.9	("sliding window" to adaptively select 10% of 100x number of points from 100x area)
BASELINE ANALYSIS						
CGFS / 79 pts / 18 nn / 1 - 64x area	**x area**	**Sec**	**Min**	**Hrs**	**Days**	**Ratio**
CGFS_base_79_pts_18_nn_01x_area	1	115.3	1.9	0.0	0.0	1.0
CGFS_base_79_pts_18_nn_02x_area	2	234.5	3.9	0.1	0.0	2.0
CGFS_base_79_pts_18_nn_04x_area	4	442.9	7.4	0.1	0.0	3.8
CGFS_base_79_pts_18_nn_08x_area	8	888.1	14.8	0.2	0.0	7.7
CGFS_base_79_pts_18_nn_16x_area	16	1710.9	28.5	0.5	0.0	14.8
CGFS_base_79_pts_18_nn_32x_area	32	3556.7	59.3	1.0	0.0	30.8
CGFS_base_79_pts_18_nn_64x_area	64	7189.3	119.8	2.0	0.1	62.3
CGFS / 1-64x pts / 18 nn / 1x area	**x pts**	**Sec**	**Min**	**Hrs**	**Days**	**Ratio**
CGFS_base_01x_pts_18_nn_1x_area	1	115.3	1.9	0.0	0.0	1.0
CGFS_base_02x_pts_18_nn_1x_area	2	118.6	2.0	0.0	0.0	1.0
CGFS_base_04x_pts_18_nn_1x_area	4	117.4	2.0	0.0	0.0	1.0
CGFS_base_08x_pts_18_nn_1x_area	8	125.4	2.1	0.0	0.0	1.1
CGFS_base_16x_pts_18_nn_1x_area	16	151.5	2.5	0.0	0.0	1.3
CGFS_base_32x_pts_18_nn_1x_area	32	214.9	3.6	0.1	0.0	1.8
CGFS_base_64x_pts_18_nn_1x_area	64	340.3	5.7	0.1	0.0	2.7
CGFS / 1024 pts / 4-1024 nn / 1x area	**x nn**	**Sec**	**Min**	**Hrs**	**Days**	**Ratio**
CGFS_base_1024_pts_0004_nn_1x_area	4	18.0	0.3	0.0	0.0	1.0
CGFS_base_1024_pts_0008_nn_1x_area	8	32.1	0.5	0.0	0.0	1.8
CGFS_base_1024_pts_0016_nn_1x_area	16	94.3	1.6	0.0	0.0	5.2
CGFS_base_1024_pts_0032_nn_1x_area	32	433.7	7.2	0.1	0.0	24.1
CGFS_base_1024_pts_0064_nn_1x_area	64	2493.5	41.6	0.7	0.0	138.4
CGFS_base_1024_pts_0128_nn_1x_area (ext.)	128	15729.6	262.2	4.4	0.2	872.9
CGFS_base_1024_pts_0256_nn_1x_area (est.)	256	98986.0	1649.8	27.5	1.1	3079.8
CGFS_base_1024_pts_0512_nn_1x_area (est.)	512	623617.0	10393.6	173.2	7.2	6614.5
CGFS_base_1024_pts_0790_nn_1x_area (est.)	*790*	*2276027.0*	*37933.8*	*632.2*	*26.3*	*5248.5*
CGFS_base_1024_pts_1024_nn_1x_area (est.)	1024	3928788.0	65479.8	1091.3	45.5	1575.6
RMNP / 1024 pts / 4-1024 nn / 1x area	**x nn**	**Sec**	**Min**	**Hrs**	**Days**	**Ratio**
RMNP_base_1024_pts_0008_nn_1x_area	9	271.0	4.5	0.1	0.0	1.0
RMNP_base_1024_pts_0016_nn_1x_area	18	443.0	7.4	0.1	0.0	1.6
RMNP_base_1024_pts_0032_nn_1x_area	36	1266.0	21.1	0.4	0.0	2.9
RMNP_base_1024_pts_0064_nn_1x_area	72	5937.0	99.0	1.6	0.1	4.7
RMNP_base_1024_pts_0128_nn_1x_area (ext.)	128	23625.7	393.8	6.6	0.3	4.0
RMNP_base_1024_pts_0256_nn_1x_area (est.)	256	138360.4	2306.0	38.4	1.6	5.9
RMNP_base_1024_pts_0512_nn_1x_area (est.)	512	810286.1	13504.8	225.1	9.4	5.9
RMNP_base_1024_pts_1024_nn_1x_area (est.)	1024	4745313.3	79088.6	1318.1	54.9	5.9
RMNP_base_1180_pts_1180_nn_1x_area (est.)	*1180*	6812384.0	*113539.7*	*1892.3*	*78.8*	1.4

C Wrapper Script to Automate Baseline CGFS Run

```
#!/bin/tcsh
#
# Script to drive the baseline Splus analysis and kriging for CGFS data.
# The kriging can be in Fortran or within Splus.
#
# Version 1.0
# July 12, 2002
#
# Check to see that the command line argument is correct
#
#
if ( ${#argv} != 3 ) then
    echo "Usage: run_baseline splus|fortran krigsize num-nearest-neighbors"
    exit(1)
endif
#
setenv mode $1
setenv krigsize $2
setenv nn $3
#
if ( ($mode != 'splus') && ($mode != 'fortran') ) then
    echo "Usage: run_baseline splus|fortran"
    exit(1)
endif
#
# Define some variables to help simplify some path expressions
#
setenv PD ./point-data
setenv ID ./image-data
setenv SD ./splus-scripts
setenv UD ./utilities
#
# Delete results of previous run
#
if ( -f $SD/sedfile ) /bin/rm -f $SD/sedfile
if ( -f $SD/splus_script ) /bin/rm -f $SD/splus_script
if ( -f $PD/cerrotp.asc ) /bin/rm -f $PD/cerrotp.asc
if ( -f $ID/cerrotp.krg ) /bin/rm -f $ID/cerrotp.krg*
if ( -f $ID/cerrotp.std ) /bin/rm -f $ID/cerrotp.std*
if ( -f $ID/cerrotp )     /bin/rm -f $ID/cerrotp*
#
# Now run Splus with the driver script
```

```
# The first calling argument selects 'splus' or 'fortran' kriging.
# The files to create the appropriate driver script for each option
#    are contained in the splus-scripts subdirectory.
# The files 'fortran_template' and 'splus_template' contain the bulk of
#    the necessary Splus commands.  A simple 'sed' script is used to
#    convert these templates into the actual driver script 'splus_script'
#    by inserting the requested output size and number of nearest numbers.
#
echo 's/PARAMETERS/'$nn','$krigsize'/' > $SD/sedfile
   sed -f $SD/sedfile < $SD/$mode'_template' > $SD/splus_script
#
echo ' '
echo 'Starting the Splus analysis at ' `date`
echo ' '
#
time /usr/local/bin/Splus << EOF_SPLUS
source("splus-scripts/splus_script")
q()
EOF_SPLUS
echo ' '
echo 'Finished the Splus analysis at ' `date`
echo ' '
#
# If requested, krig the residuals in cerrotp.asc with Fortran program.
#
if ( $mode == 'fortran' ) then
#
echo ' '
echo 'Starting the Fortran kriging at ' `date`
echo ' '
#
# Run the Fortran kriging program
#
time ./kriging/krigfor << EOF_KRIGING
$PD/cerrotp.asc
$ID/cerrotp.krgtmp
$ID/cerrotp.stdtmp
EOF_KRIGING
#
echo ' '
echo 'Finished the Fortran kriging at ' `date`
echo ' '
#
endif     #fortran kriging
#
# Now convert the standard ASCII formatted file containing the kriged
```

```
# residuals to a binary raster format and delete the intermediate file.
#
$UD/tobin $ID/cerrotp.krgtmp $ID/cerrotp.krg
$UD/tobin $ID/cerrotp.stdtmp $ID/cerrotp.std
/bin/rm -f $ID/cerrotp.krgtmp
/bin/rm -f $ID/cerrotp.stdtmp
#
# Begin IDL processing to apply the OLS fit results to the input image data,
#    adding in the kriged residuals, to create the output predictive map
#
#the following allows for S-plus kriging to be run for a smaller output area
setenv nsamples 652
setenv nlines 715
setenv krg_nsamples 652
setenv krg_nlines 715
#
if ( $krigsize == 256 ) then
   setenv krg_nsamples 257
   setenv krg_nlines 281
endif
#
cd $ID
time idl << EOF_IDL
;
; Begin by reading the maps of the significant independent variables.
;
;read the input elevation map
elv=fltarr($nsamples,$nlines) & openr,1,'./cerro-elv' & readu,1,elv & close,1
;
;read the input slope map
slp=fltarr($nsamples,$nlines) & openr,1,'./cerro-slp' & readu,1,slp & close,1
;
;read the input tassel cap coefficient #1 map
taslc1=fltarr($nsamples,$nlines) & openr,1,'./taslc1' & readu,1,taslc1 &
close,1
;
;read the input tndvi map
tndvi =fltarr($nsamples,$nlines) & openr,1,'./tndvi' & readu,1,tndvi &
close,1
;
;read the input study area definition map (or mask)
mask =bytarr($nsamples,$nlines) & openr,1,'./studyarea' & readu,1,mask &
close,1
;
;read the kriged residuals
tpkrg=fltarr($krg_nsamples,$krg_nlines) & openr,1,'./cerrotp.krg' &
```

```
readu,1,tpkrg & close,1
;
;read the estimated uncertainty from the kriging process
tpstd=fltarr($krg_nsamples,$krg_nlines) & openr,1,'./cerrotp.std' &
readu,1,tpstd & close,1
;
; Apply the OLS fit to the significant indepent variables, add the
; kriged residuals, and apply the study area mask to create the
; final output total plant map.
;   (the kriged residuals are resampled to a larger size if the Splus
;    kriging was selected.  The resampling is via cubic convolution.)
;
cerrotp = ( 94.85183 - 0.01172346 * elv - 0.3414609 * slp + 4.050923 *
tndvi - 0.1285297 * taslc1 + congrid(tpkrg,$nsamples,$nlines,cubic=-0.5)
) * float(mask)
;
;set negative values to zero and write to disk
;
cerrotp(where(cerrotp le 0.0)) = 0.0
openw,1,'./cerrotp' & writeu,1,cerrotp & close,1
;
;create illustrative JPEG and PNG file of the output total plant map
;
set_plot,'z'
device,set_resolution=[$nsamples,$nlines]
tvscl,cerrotp
write_jpeg,'cerrotp.jpg',tvrd()
write_png,'cerrotp.png',tvrd()
;
;all done with IDL processing
;
exit
;
EOF_IDL
#
# Final postprocessing steps to create ENVI header files
#
# Make ENVI header for estimated total plants file (cerrotp)
cat > ./cerrotp.hdr << EOF_HDR1
ENVI
description = {
   Cerro Grande Total Plant Prediction}
samples = $nsamples
lines   = $nlines
bands   = 1
header offset = 0
```

```
file type = ENVI Standard
data type = 4
interleave = bsq
sensor type = Unknown
byte order = 0
band names = {Total Plants}
EOF_HDR1
#
# Make ENVI header for cerrotp.krg file
cat > ./cerrotp.krg.hdr << EOF_HDR2
ENVI
description = {
   Kriged Residuals}
samples = $krg_nsamples
lines   = $krg_nlines
bands   = 1
header offset = 0
file type = ENVI Standard
data type = 4
interleave = bsq
sensor type = Unknown
byte order = 0
band names = {Kriged Residuals}
EOF_HDR2
#
# Make ENVI header for cerrotp.std file
cat > ./cerrotp.std.hdr << EOF_HDR3
ENVI
description = {
   Uncertainty of Kriged Residuals}
samples = $krg_nsamples
lines   = $krg_nlines
bands   = 1
header offset = 0
file type = ENVI Standard
data type = 4
interleave = bsq
sensor type = Unknown
byte order = 0
band names = {Uncertainty of Kriged Residuals}
EOF_HDR3
#
# All done...
#
echo 'All done at ' `date`
#
```

D Serial FORTRAN Kriging Code

```
      program krigfor
c-----------------------------------------------------------------------
c  Program to perform ordinary kriging.  Measurements of a particular
c  value (e.g. plant diversity) at discrete locations are used to
c  determine continuous estimates of the parameter across a region.
c
c  The program reads an ASCII input file containing values measured at
c  distinct spatial locations.  The input file contains a header that
c  controls the details of the kriging process.
c
c  The program writes an ASCII output file containing the estimates of
c  the value throughout the region, and optionally writes an ASCII
c  output file containing the standard errors due to the kriging.
c
c  Compilation Parameters:
c     maximum number of observation (nobs) = 2000
c     maximum number of nearest neighbors (maxn) = 200
c     maximum number of columns in kriged surface (cols) = 2000
c     unlimited number of rows
c
c  Version 1.0.
c  July 15, 2002
c
c  Written by Robin Reich, Colorado State University.
c  Documented by Jeff Pedelty, NASA's Goddard Space Flight Center
c-----------------------------------------------------------------------
      implicit none
c
      integer nobs,maxn,cols
      parameter(nobs=2000,maxn=200,cols=2000)
      integer nrow,ncol,nn,n,indx(nobs),id(nobs)
      integer i,j,k,ll
c
      real x,y,cell,xl,yl,z(nobs,3),nugget,sill,range,dst(nobs),
     *     temp(maxn,3),cova(maxn,maxn),covb(maxn),tmpcov(maxn),
     *     zhat(cols),d,sehat(cols)
      real krgmin,krgmax,stdmin,stdmax
c
      character*32 filein,fileout,filese
      character*3  model
      character*4  se
c
c-----------------------------------------------------------------------
c
```

```
c Initial values for overall minimum and maximum of the output surfaces.
c
       krgmin =  1.e5
       krgmax = -1.e5
       stdmin =  krgmin
       stdmax =  krgmax
c
c Accept the name of the input and output data files from standard input
c
       write(*,*) 'Enter the name of the input file'
       read(*,15) filein
   15  format(a32)
       write(*,*) 'Enter the name of the krig output file'
       read(*,15) fileout
c
c Open the input data file
c
       open(11,file=filein)
c
c Read the control parameters from the input file
c
c nrow   = Number of rows in output kriged surface
c ncol   = Number of columns in output kriged surface
c
       read(11,*) nrow
       read(11,*) ncol
c
c cell   = Cell size of output surface
c xl     = X coodinate of the lower left corner of output surface
c yl     = Y coodinate of the lower left corner of output surface
c
       read(11,*) cell
       read(11,*) xl
       read(11,*) yl
c
c nn     = Number of nearest neighbor data points to use in kriging
c n      = Number of data points in the input data file
c
       read(11,*) nn
       read(11,*) n
c
c nugget = Nugget of the variogram model
c range  = Range of the variogram model
c sill   = Sill of the variogram model
c model  = Functional form of the variogram model (e.g. gaussian)
c
```

```
       read(11,*) nugget
       read(11,*) range
       read(11,*) sill
       read(11,*) model
c
c se     = Logical flag to control whether to write the errors
c
       read(11,*) se
c
c Now read the data values themselves into variable z
c   z(n,1) = X location   |
c   z(n,2) = Y location   | for data point 'n'
c   z(n,3) = Value        |
c
       do 20 i=1,n
          read(11,*) (z(i,j),j=1,3)
    20 continue
c
c Done with the input file, so now close the file
c
       close(11)
c
c Open the output data file that will contain the kriged results
c
       open(12,file=fileout,form='formatted',access='sequential')
c
c Write out a simple header to describe the output kriged surface
c   The header elements are described above
c
       write(12,*) 'NCOLS',ncol
       write(12,*) 'NROWS',nrow
       write(12,*) 'XLLCORNER',xl
       write(12,*) 'YLLCORNER',yl
       write(12,*) 'CELLSIZE',cell
c
c If the standard errors were requested, then ask for the name of
c the error file, open the file, and write the header elements.
c
       if(se .eq. 'TRUE') then
c
          write(*,*) 'Enter the name of the standard error output file'
          read(*,15) filese
          open(13,file=filese,form='formatted',access='sequential')
          write(13,*) 'NCOLS',ncol
          write(13,*) 'NROWS',nrow
          write(13,*) 'XLLCORNER',xl
```

```
          write(13,*) 'YLLCORNER',yl
          write(13,*) 'CELLSIZE',cell
c
       endif
c
c Begin the kriging process itself, which is a double do loop over the
c rows and columns of the output surface
c
c Calculate the Y value for the first row to be processed
c
       y=(nrow-1)*cell + yl
c
c Start loop over the output rows, indexed by 'i'
c
       do 200 i=1,nrow
c
c Set the X value for the first column to be processed
c
       x=xl
c
c Start loop over the output columns, indexed by 'j'
c
       do 180 j=1,ncol
c
c Create index array that will be used to sort the input data
c
          do 21 k=1,n
             indx(k)=k
    21    continue
c
c Compute the Euclidean distance of this point (i,j) to each input datum
c
          do 50 k=1,n
             dst(k)=sqrt((x-z(k,1))**2+(y-z(k,2))**2)
    50    continue
c
c Sort the distance vector to find the 'nn' nearest neighbors
c The 'indx' array will contain indices to the actual sorted data
c
          call sort2(nn,n,dst,indx,nobs)
c
c Create 'temp' array to contain the 'nn' nearest neighbor data points,
c sorted by distance
c
          do 75 k=1,nn
             temp(k,1)=z(indx(k),1)
```

```
            temp(k,2)=z(indx(k),2)
            temp(k,3)=z(indx(k),3)
    75    continue
c
c Compute distance matrix for these 'nn' data points.
c Place in array 'cova' to save on storage
c Matrix contains the Euclidean distances between the data points
c
          do 100 k=1,nn
             do 99 ll=1,nn
                cova(k,ll)=sqrt((temp(k,1)-temp(ll,1))**2+
      *         (temp(k,2)-temp(ll,2))**2)
    99       continue
   100    continue
c
c Compute distance vector between these 'nn' data points and the
c output point (i,j) whose value is being estimated
c
          do 120 k=1,nn
              covb(k)=sqrt((x-temp(k,1))**2+(y-temp(k,2))**2)
   120    continue
c
c The 'cov' subroutine converts the 'cova' and 'covb' arrays
c from distance arrays to covariance arrays by applying the
c variogram model
c
          call cov(nn,cova,covb,nugget,sill,range,model,maxn)
c
c Copy the 'covb' array to 'tmpcov' for safe keeping, as 'covb'
c will be overwritten in next steps
c
          do 130 k=1,nn+1
             tmpcov(k)=covb(k)
   130    continue
c
c The subroutines 'ludcmp' and 'lubksb' perform LU decomposition
c with backsubstitution to invert the 'cova' matrix and multiply
c the inverse by the 'covb' vector.  Resulting vector is returned
c in 'covb', and contains the weights needed to form a linear
c combination of the input data values.
c
          call ludcmp(cova,nn+1,maxn,id,d)
          call lubksb(cova,nn+1,maxn,id,covb)
c
c Calculate dot product of the input data values and the 'covb' array
c This is the appropriate weighted average to estimate the value at
```

```
c this point (i,j).
c
          zhat(j)=0.
          sehat(j)=0.
          do 150 ll=1,nn
            zhat(j)=zhat(j)+temp(ll,3)*covb(ll)
   150    continue
c
c Determine if this value is a new extremum
c
          krgmin = amin1 (krgmin, zhat(j))
          krgmax = amax1 (krgmax, zhat(j))
c
c Estimate the error at this point, if requested
c
          if (se .eq. 'TRUE') then
c
             do 160 ll=1,nn+1
                sehat(j)= sehat(j)+tmpcov(ll)*covb(ll)
   160       continue
c
c Evaluate if this error estimate is a new extremum
c
             sehat(j)=sqrt(sill-sehat(j))
             stdmin = amin1 (stdmin, sehat(j))
             stdmax = amax1 (stdmax, sehat(j))
c
          endif
c
c Done with this (i,j) point, so now update the X coordinate for next
c
          x=x+cell
c
   180 continue
c
c Done with an entire row of the output surface, so write it to output
c file, along with the errors, if requested.
c
          write(12,*) (zhat(ll),ll=1,ncol)
          if(se .eq. 'TRUE') then
             write(13,*) (sehat(ll),ll=1,ncol)
          endif
c
c Update the Y coordinate for the next row
c
          y=y-cell
```

```
c
   200 continue
c
c Write the overall minimum and maximum value of the kriged surface,
c and, if requested, the minimum and maximum of the error surface
c
       print *, 'kriged min, max = ', krgmin, krgmax
       if (se .eq. 'TRUE') print *, 'error min, max = ', stdmin, stdmax
c
c All done, so finally close the output files
c
       close(12)
       if (se .eq. 'TRUE') close(13)
c
       end
```

E Cerro Grande Background Paper

ACCEPTED FOR PUBLICATION – Proceeding of the 22nd Tall Timbers Fire Ecology Conference: Fire in Temperate, Boreal, and Mountain Ecosystems.

July 13, 2002
Mohammed A. Kalkhan
Natural Resource Ecology Laboratory,
Colorado State University, Fort Collins, CO 80523-1499, USA,
(970) 491-5262, FAX (970) 491-1965, mohammed@nrel.colostate.edu

INTEGRATION OF SPATIAL INFORMATION AND SPATIAL STATISTICS: A CASE STUDY OF INVASIVE PLANTS AND WILDFIRE ON THE CERRO GRANDE FIRE, LOS ALAMOS, NEW MEXICO, USA

Mohammed A. Kalkhan, Erik J. Martinson, Philip N. Omi,
Thomas J. Stohlgren, Geneva W. Chong, and Molly A. Hunter
Natural Resource Ecology Laboratory (NESB-A244), Colorado State University, Fort Collins, CO 80523-1499, USA.

ABSTRACT
Investigating spatial relationships among fuels, wildfire severity, and post-fire invasion by exotic plant species through linkage of multi-phase sampling design and multi-scale nested sampling field plots, pre- and post-fire, can be accomplished by integration of spatial information using spatial statistical models. This technique provides useful information and the tools for describing ecological and environmental characteristics including landscape-scale fire regimes, invasive plants, and hot spots of diversity (native and non-native plants) for the Cerro Grande fire site, Los Alamos, NM, USA. To predict the distribution, presence, and patterns of native and exotic species, we used modeling of large-scale and small-scale variability by integrating field data and spatial information (eight bands of Landsat TM Data, six derived vegetation indices and six bands of tasseled cap transformations, elevation, slope, aspect) and spatial statistics. We present the results of trend surface models that describe the large-scale spatial variability using stepwise multiple regressions based on the Ordinary Least Squares (*OLS*) method. Models with small variance were selected. In addition, the residuals from the trend surface model based on the *OLS* estimates were modeled using ordinary kriging for modeling small-scale variability based on a Gaussian semi-variogram. The final surfaces were obtained by combining two models (the trend surface based on the *OLS* and the kriging surface of residuals). All models were selected based on the lowest values of standard errors, AICC statistics, and high R^2. For large-scale spatial variability models using the *OLS* procedure, R^2 values ranged from 10.04% to 58.6% and all variables were significant at á < 0.05 level. When adding the kriging model with the OLS model, R^2 values ranged from 60% to 84%.

Keywords: invasive plants, fire severity, multi-phase design, multi- scale sampling, spatial information, spatial statistics, trend surface, OLS, variogram, kriging

Citation: Mohammed A. Kalkhan, Erik J. Martinson, Philip N. Omi,
Thomas J. Stohlgren, Geneva W. Chong, and Molly A. Hunter. Integration of Spatial Information and Spatial Statistics: A Case Study of Invasive Plants and Wildfire on the Cerro Grande Fire, Los Alamos, New Mexico, USA. Pages 000-000 *in* R. T. Engstorm and W. J. de Groot (eds.) Proceeding of the 22nd Tall Timbers Fire Ecology Conference: Fire in Temperate, Boreal, and Mountain Ecosystems. Tall Timbers Research Station, Tallahassee, FL

INTRODUCTION

Synergistic interactions and positive feedbacks among fuels, extreme wildfire behavior, and exotic species invasions are increasingly recognized as major threats to the structure and function of natural ecosystems (Mack and D'Antonio 1998). We are currently investigating spatial relationships among fuels, wildfire severity, post-fire invasion by exotic plant species, and other ecological – environmental characteristics through the linkage of multi-phase design (Fig. 1), multi-scale field plots (Modified-Whittaker, Stohlgren et al. 1995, 1998, Fig. 2), and pre- and post-fire remote sensing imagery using spatial models (Kalkhan et al. 1998, Kalkhan and Stohlgren 2000, Kalkhan et. al. 2001). The integration of spatial information (remote sensing data, Geographic Information Systems [GIS]) using spatial statistics provides useful tools for assessing landscape-scale structure of forest and rangelands (Kalkhan et. al. 2000, 2001, Chong et al. 2001). In addition, the ability to model the small-scale variability in landscape characteristics requires the generation of full-coverage maps depicting characteristics measured in the field (Gown et al.1994). Gown et al. (1994) point out that, while many spatial datasets describing land characteristics have proven reliable for macro-scale ecological monitoring, these relatively coarse-scale data fall short in providing the precision required by more refined ecosystem resource models.

Reich et al. (1999) described a model based on the process using stepwise regression, trend surface analysis of geographical variables (e.g., elevation, slope, and aspect), and measures of local taxa to evaluate large-scale spatial variability. This model was used in this study and is defined as:

$$\Phi_0 = \sum_{i+}^{p} \sum_{j \le p}^{p} \beta_{ij}\, x_{10}^{i}\, x_{20}^{j} + \sum_{K=1}^{q} \gamma_k\, y_{k0} + \eta_0 \qquad \textbf{(1)}$$

where β_{ij} are the regression coefficients associated with the trend surface component of the model, γ_k are the regression coefficients associated with the q auxiliary variables, y_{k0}, are available as a coverage in the GIS data base, and η_0 is the error term which may or may not be spatially correlated with its neighbors (Kallas 1997, Metzger 1997).

Spatial statistics and spatial information provide a means to develop spatial models that can be used to correlate coarse-scale geographical data with field measurements of biotic variables. Here we present our spatial modeling process and preliminary predictive models of native and exotic plant distributions for the 2000 Cerro Grande fire, Los Alamos, NM.

Our research program objectives included the interpolation of plot-level information to the landscape-scale with generalized predictive spatial statistical models derived from remotely sensed data, GIS, and field data to allow broad examination and conclusions regarding the interactions among fuels, wildfire, and exotic plants. The uniqueness of this approach is to allow using the combination of multi-phase sampling design (i.e., double sampling, Kalkhan et al. 1998, Fig. 1) and multi-scale nested plot designs (Modified-Whittaker, Stohlgren et al. 1995, 1998). The main plot dimension is 20 m x 50 m (1000 m^2) with ten 0.5 m x 2 m (1 m^2) subplots, two 2 m x 5 m (10 m^2) subplots in opposite corners, and a 5 m x 20 m (100 m^2) subplot in the plot center (Fig. 2). Both approaches allow us to perform intensive unbiased sampling surveys at certain plot levels which can help to reduce the cost of sampling surveys and improve the efficiency of sampling design .The specific objective of this paper is to develop a predictive spatial statistical model for describing large- and small-scale variability of plant species richness (native and exotic species) on the Cerro Grande fire, Los Alamos, New Mexico, USA.

STUDY SITE SELECTION

The Cerro Grande fire site is located near Los Alamos, New Mexico with elevation range from 1932m to 3200m. The fire site was well-suited for our study because it included multiple fuel types, exhibited a wide range of burn severities, and involved pre-fire fuel treatments. In addition, existing digital spatial information was abundant and available, and there was potential for cooperation with other research groups that have complementary interests.

We completed field sampling for our study site in August 2001. The Cerro Grande fire began as a prescribed fuel treatment by Bandelier National Monument, Los Alamos, NM on 4 May 2000. The fire escaped control and was declared a wildfire on 5 May 2000. The fire was contained on 24 May after burning about 19,300 ha of lands managed by seven different agencies, including the town of Los Alamos, NM. However, 60% of the fire area burned 10-11 May, 2000, and 60% of the fire was on the

Española District of the Santa Fe National Forest (Burned Area Emergency Rehabilitation [BAER] Team 2000). Initial remotely sensed estimates of burn severity were classified as high (35%), moderate (9%), and low (56%). Elevations in sampled areas ranged from 2,000 m to 3,000 m and included pinyon-juniper woodlands, ponderosa pine forests, and mixed-conifer forests.

METHODS

Sampling Design

We employed a stratified random sampling design to locate 66 multi-scale nested plots (Modified-Whittaker, Stohlgren et al. 1995, 1998, Fig. 2) within areas burned 10-11 May 2000 in the Santa Fe National Forest, and an additional 13 unburned plots within 300 m of the fire perimeter. Burned area strata included vegetation type (pinyon-juniper woodland, ponderosa pine forest, and mixed conifer forest), BAER fire severity classification (high, low, moderate), aspect (north, south), and pre-fire fuel treatment (untreated, prescribed burn, thin only, thin followed by prescribed burn). Unburned strata included aspect (north, south) and elevation (<2500m, >2500m). At least three plots were randomly located in each stratum.

Data Analysis

Data collected from each plot included measurements related to pre-fire stand condition, refined estimates of fire severity, plant species cover and richness, and measurements related to post-fire fuel flammability. For the vegetation data, we used the Modified-Whittaker multi-scale nested plots design (Fig. 2). The global positioning system was used to document the locations of the plots and incorporate the field data directly into the GIS. Five soil samples (10-20 cm depth) were taken and pooled from each 20 m x 50 m vegetation plot. These five samples were located in each of the corners of each Modified-Whittaker plot, as well as in the plot center. Samples were used for total carbon (C), nitrogen (N), and soil texture analyses. Data used in modeling included eight bands of Landsat TM Data, six different vegetation indices, six bands of transformed tasseled cap indices (using ERDAS-IMAGINE 8.4**,** ERDAS 2000), topographic derived data (elevation, slope, aspect; ARCINFO version 7.4**,** ESRI 2000), and vegetation data (total number of plant species, number of native plant species, number of exotic plant species, and percent cover for total, native, and exotic species). All spatial information from remotely sensed data and GIS layers were converted to a grid using ARCINFO (ESRI 2000, version 7.4), and a program written in AML (*ARC MACRO LANGUAGE,* ESRI 2000) was used to extract the 79 data points (field plot locations) with respect to their UTM-X and Y coordinates within the study area. All data were then used for the development of the spatial models using S-plus software (MathSoft 2000). The soil data were not used in this paper, but it will be used in future research papers to include sub-plot level information for both soil and vegetation.

Spatial Analysis

In this paper we used the same approach by Kalkhan and Stohlgren (2000) by using the cross-correlation statistic to test the null hypothesis of no spatial cross-correlation among all pairwise combinations of vegetation variables and topographic characteristics (Table 1). In calculating the cross correlation-statistic (I_{YZ}), the inverse distance between sample plots was used as a weighting factor to give more weight to values in the closest sample plots and less to those in plots that are farthest away. The null hypotheses of no spatial cross-correlation were rejected when the *P*-value associated with the test statistic was less than 0.05. *Moran's I*, which is a special case of the cross-correlation statistic I_{YZ} (Czaplewski and Reich 1993), was used to calculate the spatial auto-correlation associated with each of the variables used in this study (Table I). Cliff and Ord (1981) showed that I_{YZ} ranges from –1 to +1, although it can exceed these limits with certain types of spatial matrices. Data distributions that were strongly skewed were transformed prior to analysis. Aspect data were transformed using the absolute value from due south (180^{o}; high solar radiation, Kalkhan and Stohlgren 2000).

Spatial Modeling

Stepwise multiple regression analysis was used first to identify the best linear combination of independent variables. It also allows us to explore the variation in predicting total, exotic, and native plant species richness as a function of the eight TM bands, six derived vegetation indices, six tasseled cap

transformation indices, slope, aspect, and elevation. The selected independent variables were used in the Ordinary Least Square (*OLS*) procedure to describe large-scale variability estimates.

OLS estimators were used to fit the model if the variable of interest had a linear relationship with the geographical coordinates of the sample plots, the digital number (DN) value of any of the Landsat TM bands, and the topographic data. In addition, the least squares method fits a continuous, univariate response as a linear function of the predicted variable. This trend surface model represented continuous first order spatial variation. Akaike's Information Criteria "AIC" (Brockwell and Davis 1991, Akaike 1997) was used as a guide in selecting the number of model parameters to include in the regression model where:

$$AIC = -2 (\text{max log likelihood}) + 2 (\text{number of parameters}) \quad \textbf{(2)}$$

When using maximum likelihood as a criterion for selecting between models of different orders, there is the possibility of finding another model with equal or greater likelihood by increasing the number of parameters (Metzger 1997). Therefore, the AIC allows for a penalty for each increase in the number of parameters. Using this criterion, a model with a smaller AIC was considered to have a better fit. While the model was kept as simplistic as possible, a more complex model could be used if the situation warrants it. In this paper, we used the AICC which is a modification model of AIC (Reich et al. 1999).

In the next stage of the model building process, the residuals from the trend surface models were analyzed for spatial dependencies. This was accomplished using spatial auto-correlation and cross-correlation statistics. If the residuals were cross-correlated with other variables, we could use co-kriging to interpolate the residuals. However, if the residuals were not cross-correlated, we used ordinary kriging. Finally, the weights associated with the kriging and co-kriging models were estimated as a function of the spatial continuity of the data (Isaaks and Srivastiva 1989). This estimation can be accomplished using a sample variogram to describe spatial continuity. With spatial data, the variation of the samples generally changes with distance. In other words, the variogram is a measure of how the variance changes with distance. The variogram and cross-variogram models used in this analysis were considered "basic" models, meaning they are simple and isotropic (Reich et al. 1999). They include Gaussian, spherical, and exponential models (see Isaaks and Srivastiva 1989). Prior to estimating the sample variogram and cross-variogram, the data were rescaled by dividing the individual variables and the residuals by their respective maximum values. This was necessary to maintain numerical stability (Isaaks and Srivastiva 1989) by eliminating any differences in the magnitude of the variables without altering the solution. Although this was not necessary for kriging, it was important in co-kriging (Isaaks and Srivastiva 1989, Metzger 1997).

RESULTS

We used 79 data points (based on Modified-Whittaker nested plots of 1000 m^2) to represent different variables that were extracted from Landsat TM data (eight bands, six vegetation indices, and six bands of tasseled cap transformation indices), topographic data (elevation, slope, and absolute aspect), and vegetation characteristics (Table 1). Total plant species richness (hot spot of plants diversity), including species of unknown origin and taxa that could not be identified, ranged from 14 to 78 per plot. Typically, non-native species represented less than 10% of the total species at a site and about 5% of the foliar cover (Table 1).

Spatial Relationships

The preliminary results for our field data using *Moran's I* (Moran 1948, Mantel 1967*)* and the bivariate cross correlation-statistic "I_{YZ}" (Czaplewski and Reich 1993, Bonham et al. 1995) to test for spatial auto-correlation and cross-correlation with residuals suggested that, at large-scales, the probabilities of presence and absence of exotic plant species and their percent cover were spatially independent throughout the study site (Table 2). That is, the spatial relationships were not statistically significant. Native species richness was not independent (Kalkhan and Stohlgren 2000). However, these results may be different for individual plant species (Kalkhan et al. 2000). In general, large-scale patterns of species distribution were controlled by topographic factors such as elevation, aspect, and slope with complex spatial patterns. This may explain why negative spatial auto-correlation and cross correlation resulted when large-scale plots were used (Kalkhan et al. 2000). These results may have been different if individual native or exotic plant species had been used in the analysis (Kalkhan et al. 2000).

Spatial Statistical Model:

The results of modeling the large-scale and small-scale variability in predicting total, native, and exotic species richness and percent cover of exotic and native plant species within the Cerro Grande fire site are shown in Table 2. Models were developed for large-scale variability of the total number of plants (both native and exotic species) and percent plant cover (total, native, and exotic). The trend surface models identified using stepwise multiple regressions that had R^2 values ranged from 10.04% to 58.6% and all variables were significant at á < 0.05 level.

Small-scale variability models are used to examine the spatial continuity of variability and were developed using ordinary kriging based on the Gaussian semi-variogram model which was based on the AICC criteria (Table 2). Model parameters were estimated using weighted least squares (Cressie 1985). The residuals were also analyzed for spatial auto-correlation and cross-correlation (Czaplewski and Reich 1993, Reich et al. 1994) with the geographical variables (e.g., elevation, slope, other). Inverse distance weighting was used to define the spatial weights matrix. The kriging models were cross-validated to assess the variability in the prediction errors. The cross-validation included deleting one observation from the data set and predicting the deleted observation using the remaining observations (Reich et al. 1999). This process was repeated for all observations in the data set. The final models (trend surface plus the kriged residuals) had R^2 values ranging from 60% to 84%. In addition, the accuracies of the kriging models were assessed using the relative mean squared error suggested by Havesi et al. (1992).

Figs. 3 and 4 represent examples of predictive spatial statistical maps based on the trend surface model (*OLS*) and kriging (variogram) on total species richness distributions for total plant species and exotic plant species within the Cerro Grande fire site. The spatial map (e.g., number of native plants) can be used in other spatial models if the native species variable is significant. Fig. 5 is an example of the standard errors associated with predicting exotic plant species richness (map of uncertainty). The figure shows standard error values increased with distance from the sample points, as would be expected. The standard error values indicated significant utility of the map of exotic plant species richness for directing future management activities. This technique of spatial mapping provides a unique way to describe landscape-scale wildfire patterns and may contribute to better management decisions. Adding more sampling points and examining ecological relationships (e.g., between vegetation and soil) may help to improve predictive spatial statistical models and their accuracy (i.e., error reductions).

DISCUSSION

Investigating spatial relationships among fuels, wildfire severity, and post-fire invasion by exotic plant species through linkage of multi-phase sampling design and multi-scale nested sampling field plots, pre- and post-fire, can be accomplished by the integration of remotely sensed data, GIS, using spatial statistical models. This technique provided useful information and tools for describing landscape-scale patterns of plant diversity within the Cerro Grande fire site, Los Alamos, NM. Current fire behavior models such as BEHAVE (Andrews 1986) and FARSITE (Finney 1998) were used to aid in predicting fire and subsequent mapping of probable scenarios of fire spread during a given time period. The disadvantage of using these types of models is the lack of using remotely sensed data. The models utilized only forest stand parameters, fire behavior, a fuel model, and topographic (i.e., elevation, aspect, and slope) characteristics. However, using remote sensing data allows us to easily develop these layers and their characteristics. Satellite data and aerial photographs have been used to map vegetation characteristics and then assign fuel models to various vegetation classes (Kourtz 1977, Mark et al. 1995, Miller and Johnston 1985, Wilson et al. 1994). The disadvantage of this approach is that the various components of vegetation (i.e., forest structure) are not always correlated with existing vegetation characteristics because of past management activities and random disturbance in the form of individual tree or plant mortality (Reich et al. 2002). Thus, collecting intensive fuel data and vegetation measurements using unbiased multi-scale sampling within the forest landscape provide an excellent data source and input to spatial models similar to the one used in this paper. These spatial models provide unbiased estimates of the various components of forest fuels as well as estimates of the prediction variance associated with individual estimates. Also, the estimating spatial models are relatively more precise and accurate in terms of statistical components and properties than currently available fuel models, and are thus more useful to the forest decision-makers. Models covering such areas as the Cerro Grande site enable the spatial integration of fuel loading estimates to a wide range of spatial

scales, along with estimates of the level of uncertainty. Finally, these types of models can help natural resource management teams to minimize field assessment by using multi-phase sampling design and multi-scale nested plot designs.

CONCLUSIONS

The integration of remotely sensed data and GIS using spatial statistics provides useful information for describing large- and small-scale variability of landscape, as demonstrated at the Cerro Grande fire site, Los Alamos, NM. We used spatial statistical predictive models based on large-scale and small-scale variability to predict plant species richness of both native and exotic plant species (hot spots of diversity) and patterns of exotic plant invasions. Large-scale spatial variability models using multiple stepwise regressions based on the OLS method had R^2 values ranging from 10% to 59%. When adding kriging with trend surface using *OLS* estimates, R^2 values ranged from 60% to 84%. All models were significant at the á < 0.05 level. The predicted standard errors for exotic species richness (Fig. 5) are less than 40% of the mean number of exotic species per plot, even at the farthest distance from a sampled point. This indicates significant utility of the map of exotic species richness for directing management activities because the error is relatively low. This error could be reduced when soil data, for example, becomes available and could add to future predictive models.

Future research will use data (including additional variables of soil and vegetation) collected from small subplots (i.e., 1 m^2). This will improve the accuracy of model predictions as well as advance the investigations of spatial auto-correlation and cross-correlation statistical patterns in landscape-scale assessments, which are essential for the development of spatial statistical models for relations between vegetation and environmental variables (e.g., soil characteristics), fuel data, and wildfire severity at different levels. This will also help us to understand their spatial relationships with respect to remotely sensed data at different scales of plot sizes (e.g., 1 m^2, 10 m^2, 100 m^2, and 1000 m^2) and improve the spatial model, since we will be able to capture more information about landscape-scale patterns and variability.

ACKNOWLEDGMENTS

Funding for this research was provided by the Joint Fire Science Program. We thank the staff of the Natural Resource Ecology Laboratory and the Department of Forest Sciences at Colorado State University for providing invaluable support. Benjamin Chemel, Steve Desipio, Li-Ming Liang, Rod Rochambeau, Rick Shory, and Sara Simonson assisted in field data collection and provided botanical expertise. The spatial statistical library used in this paper was provided and developed by Dr. Robin M. Reich, Department of Forest Sciences, and Dr. Richard A. Davis, Department of Statistics, Colorado State University. Dr. Deborah R. Spotts provided helpful editorial comments for this paper. We thank Mr. Tom Mellon, USDA Forest Service in Albuquerque, NM for providing remote sensing and GIS data. To all we are grateful.

LITERATURE CITED

Akaike, H. 1997. On entropy maximization principal. *In*: Applications of Statistics, P.R. Krishnaiah (ed.). Pp. 27-41, North-Holland, Amsterdam.

Andrews, P.L. 1986. BEHAVE. Fire behavior prediction and fuel modeling system. Burn subsystems, USDA Forest Service, Ogden, Utah.

Bonham, C.D., R.M. Reich, and K.K. Leader. 1995. A spatial cross-correlation of *Bouteloua gracilis* with site factors. Grasslands Science 41:196-201.

Burned Area Emergency Rehabilitation Team. 2000. Cerro Grande Fire Burned Area Emergency Rehabilitation (BAER) Plan. Available online at: http://www.baerteam.org/cerroGrande/Rehabilitation%20Plan.pdf.

Brockwell, P.J. and R.A. Davis. 1991. Time series. Theory and Methods. Springer, New York, 577 p.

Chong, G.W., R.M. Reich, M.A. Kalkhan, and T.J. Stohlgren. 2001. New approaches for sampling and modeling native and exotic plant species richness. Western North American Naturalist 61:328-335.

Cliff, A.D. and J.K Ord. 1981. Spatial Processes, Models and Applications. Pion Ltd., London, England, pp. 21-45.

Cressie, N. 1985. Fitting variogram models by weighted least squares. Mathematical Geology 17:563-586.

Czaplewski, R.L. and R.M. Reich. 1993. Expected value and variance of Moran's bivariate spatial autocorrelation statistic under permutation. USDA For. Serv. Res. Paper RM-309. Fort Collins, Colorado, pp. 1-13.

ERDAS – IMAGINE. 2000. ERDAS Inc., Atlanta, GA.

ESRI. 2000. Environmental Systems Research Institute, Inc. 380 New York St., Redlands, CA 97373 USA.

Finney, M.A. 1998. FARSITE: Fire area simulator-area model. Development and evaluation. RMRS-RP-4, Ogden, Utah: USDA Forest Service, Rocky Mountain Research Station, 47 p.

Gown, S.N., R.H. Waring, D.G. Dye, and J. Yang. 1994. Ecological remote sensing at OTTER: Satellite Macroscale Observation. Ecological Applications 4:322-343.

Havesi, J.A., J.D. Istok, and A.L. Flint. 1992. Precipitation estimation in mountainous terrain using multivariate geostatistics. Part I: structural analysis. Journal of Applied Meteorology 3:661-676.

Isaaks, E.H. and R.M. Srivastiva. 1989. An introduction to applied geostatistics, Oxford University Press, New York, 561 p.

Kalkhan, M.A., T.J. Stohlgren, G.W. Chong, L.D. Schell, and R.M. Reich. 2001. A predictive spatial model of plant diversity: Integration of Remotely Sensed data, GIS, and Spatial statistics. Proceeding of the Eighth Forest Remote Sensing Application Conference (RS 2000), April 10-14, 2000, Albuquerque, New Mexico, 11 pp. CD-ROMs Publications (ISBN 1-57083-062-2).

Kalkhan, M.A., G.W. Chong, R.M. Reich, and T.J. Stohlgren. 2000. Landscape-scale assessment of mountain plant diversity: Integration of Remotely Sensed Data, GIS, and Spatial Statistics. *In*: ASPRS Annual Convention & Exposition, ASPRS Technical Papers, May 22-26, 2000, Washington, DC. ASPRS - CD -ROMs Publication (ISBN 1-57083-061-4) by Clearance Center, Inc, 222 Rosewood Drive, Danvers, MA 01923 (Adobe Acrobat Reader Format, p. 163).

Kalkhan, M.A. and T.J. Stohlgren. 2000. Using multi-scale sampling and spatial cross-correlation to investigate patterns of plant species richness. Environmental Monitoring and Assessment 64:591-605.

Kalkhan, M.A., R.M. Reich, and T.J. Stohlgren. 1998. Assessing the accuracy of Landsat Thematic Mapper map using double sampling. International Journal of Remote Sensing 19:2049-2060.

Kallas, M. 1997. Hazard rating of Armillaria root rot on the Black Hills National Forest. M.S. Thesis, Department of Forest Sciences, Colorado State University, Fort Collins, CO 80523.

Kourtz, P.H. 1977. An application of Landsat digital technology to forest fire fuel mapping. *In*: Dube, D.E. (ed.). Fire Ecology in Resource Management: A Workshop. Northern Forest Research Center, Edmonton, Alberta. Information Report NOR-X-210. Pp. 79-81.

Mack, M.C. and C.M. D'Antonio. 1998. Impacts of biological invasions on disturbance regimes. Trends in Ecology and Evolution 13:195-198.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27:209-220.

Mark, C.A., C.L. Bushey, and W. Smetanka. 1995. Fuel Model identification and mapping for fire behavior prediction in the Absaroka-Beartooth Wilderness, Montana and Wyoming. *In*: Brown J.K., Mutch, R.W., Spoon, C.W., and Wakimoto, R.H. (tech. coord.), Proceeding symposium on fire in wilderness and park management. USDA Forest Service General Technical Report INT-GTR-320, pp. 227-229.

MathSoft Inc. 2000. MathSoft Inc., Seattle, WA.

Metzger, K. 1997. Modeling small-scale spatial variability in stand structure using remote sensing and field data. M.S. Thesis, Dept. of Forest Sciences, Colorado State University, Fort Collins, CO 80523.

Miller, C.A. and D. Johnston. 1985. Comparison of fire fuel maps produced using MSS and AVHRR data. In: Proceeding of the Pecora X Symposium (Bethesda, MD: ASPRS). Pp. 305-314.

Moran, P.A.P. 1948. The interpretation of statistical maps, Royal Statistics Society, Series. B, 10: 243-351.

Reich, R.M., J.E. Lundquist, and V.A. Bravo. 2002. Spatial models for examining the effect of diseases and other disturbance on patterns of fuel loading. International Journal of Wildland Fire (In review).

Reich, R.M., C. Aguirre-Bravo, M.A. Kalkhan, and V.A. Bravo. 1999. Spatially based forest inventory for Ejido El Largo, Chihuahua, Mexico. *In*: North America Science Symposium Toward a unified framework for inventorying and monitoring forest ecosystem resources, November 1-6, 1998, Guadalajara, Jalisco, Mexico, USDA Forest Service, Rocky Mountain Research Station, Proceedings RMRS-P-12, December 1999, pp. 31-41.

Stohlgren, T.J., K. A. Bull, and Y. Otsuki. 1998. Comparison of rangeland vegetation sampling techniques in the Central Grasslands. Range Management 51:164-172.

Stohlgren, T.J., M.B. Falkner, and L.D. Schell. 1995. A modified-Whittaker nested vegetation sampling method. Vegetatio 117:113-121.

Wilson, B. A., C.F.W. Ow, M. Heathcott, D. Milne, T.M. McCaffrey, and S.E. Franklin. 1994. Landsat MSS classification of fire fuel types in Wood buffalo National Park. Global Ecology and Biogeography Letters 4:33-39.

Table 1. Summary statistics for all variables used in developing spatial statistical models for the Cerro Grande fire, Los Alamos, NM.

Variable	Minimum	Median	Mean	Maximum
Total Plants Species	14	44	51	78
Native Plants Species	8	31	40	57
Exotic Plants Species	0	4	4.1	9
Native Cover (%)	4.2	22.3	25.9	76.3
Exotic Cover (%)	0	0.6	1.3	7.9
Elevation	1972	2266	2356	3023
Slope	1.4	10.02	12.46	32.5
Absolute Aspect	5.2	80	86.9	180
TM- Band 1	60	80	81.3	116
TM- Band 2	45	65	66.3	106
TM- Band 3	38	71	73.5	131
TM- Band 4	29	48	49.9	111
TM- Band 5	43	100	98.9	168
TM- Band 6	112	188	185.1	222
TM- Band 7	26	92	92.2	169
TM- Band 8	34	47	49.2	85
Band (5/4)	63	127	133.5	191
Band (4/3)	1	1	1.038	2
Band (3/1)	85	85	88.2	170
Band (4 – 3)	22	42	54.9	184
NDVI	0	1	0.620	1
TNDVI	0	0	0.4975	115
Tassel Cap – Band 1	111	168	173.4	265
Tassel Cap – Band 2	-80	-53	-49.8	3
Tassel Cap – Band 3	-83	-41	38.7	7
Tassel Cap – Band 4	19	27	26.7	34
Tassel Cap – Band 5	-71	-39	-37.7	-12
Tassel Cap – Band 6	-20	-15	-15.2	-11

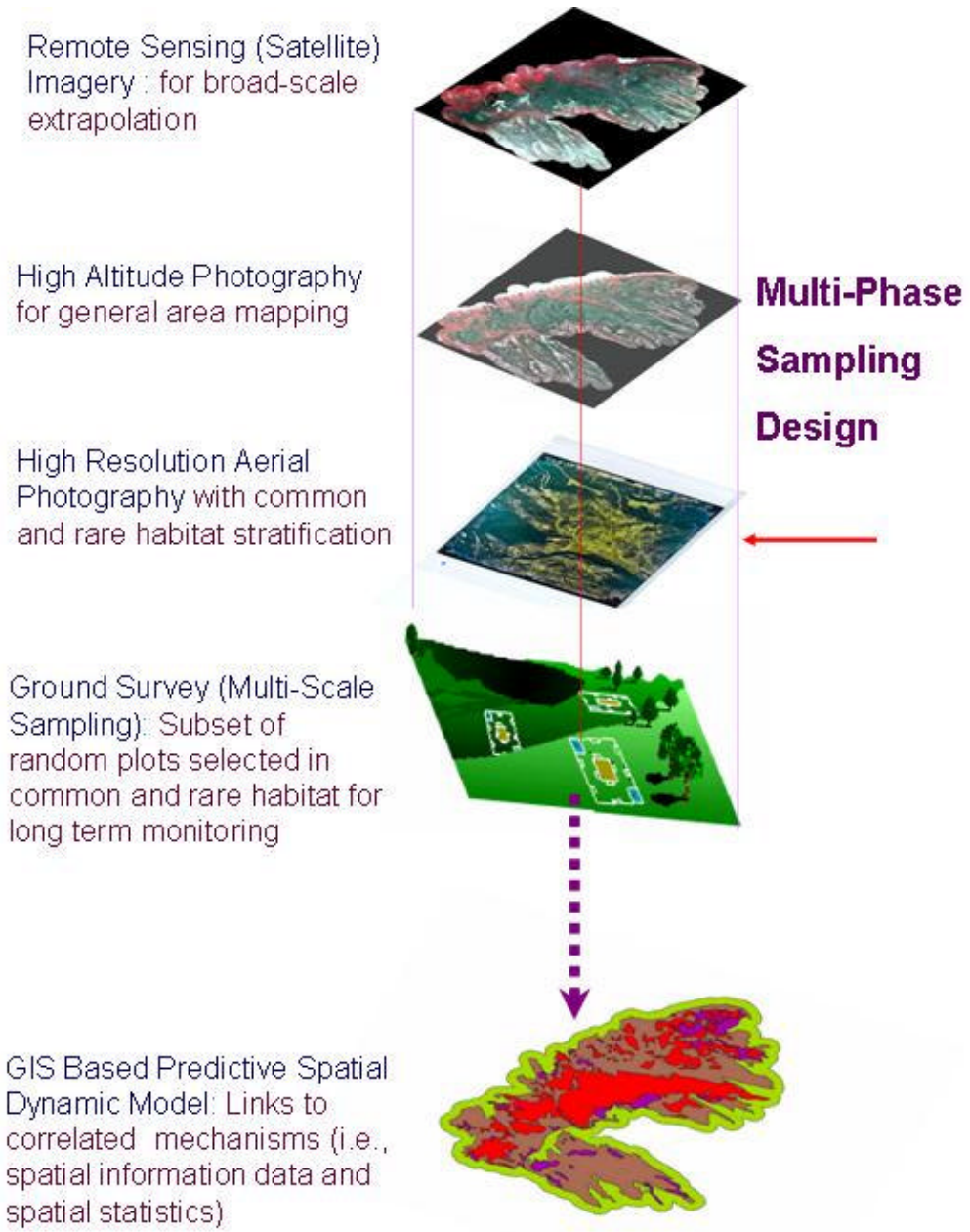
Table 2. Summary statistics for large-scale and small-scale variability models for predicting total, native, and exotic plant species richness and their precent cover within the Cerro Grande fire, Los Alamos, NM.

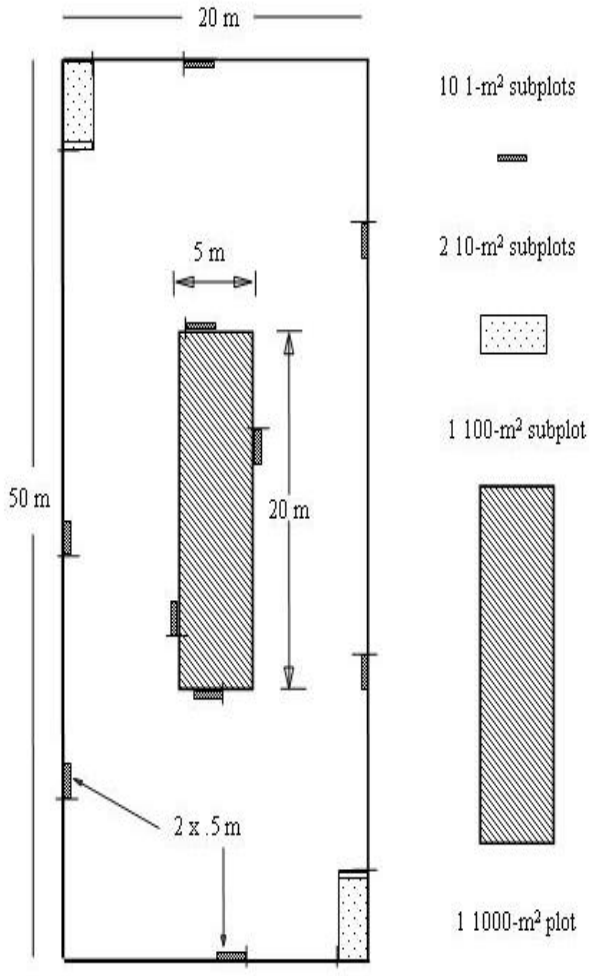
Large-scale Variability (*OLS* Model)				Large-scale and Small- scale Variability (*OLS* and Kriging-variogram Model)			
Variables	R^2 (%)	S.E.	AICC	Model	R^2 (%)	S.E.	
Total Plant Species	14.1	11.1	610.3	Gaussian	63.9	7.0	
Native Plant Species	43.7	8.6	571.6	Gaussian	60.0	7.0	
Exotic Plant Species	58.2	1.6	309.5	Gaussian	60.9	1.5	
Probability of Exotic Species	58.6	1.97	342.1	No Spatial Auto-Correlation with Residuals			
Total Plant Cover (%)	43.6	13.3	639.6	Gaussian	81.6	7.3	
Native Plant Cover (%)	46.2	13.3	639.9	Gaussian	84.4	6.9	
Exotic Plant Cover (%)	10.1	0.5	125.2	No Spatial Auto-Correlation with Residuals			

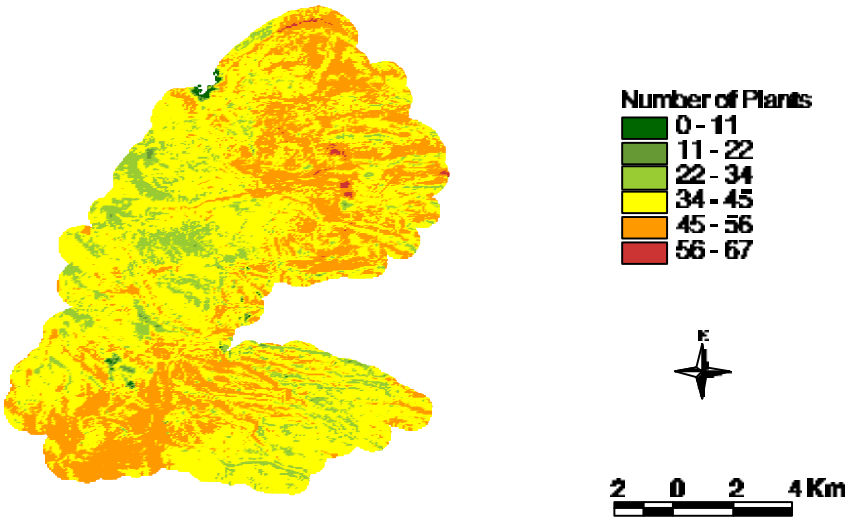
The *p* - values were significant at á < 0.05 level for the *OLS* models and significant at á < 0.01 for the variogram models.

S. E. = Standard Errors.

List of Figures


Remote Sensing (Satellite) Imagery : for broad-scale extrapolation
High Altitude Photography for general area mapping
High Resolution Aerial Photography with common and rare habitat stratification
Ground Survey (Multi-Scale Sampling): Subset of random plots selected in common and rare habitat for long term monitoring
GIS Based Predictive Spatial Dynamic Model: Links to correlated mechanisms (i.e., spatial information data and spatial statistics)
Multi-Phase Sampling Design


20 m
50 m
5 m
20 m
2 x .5 m
10 1-m² subplots
2 10-m² subplots
1 100-m² subplot
1 1000-m² plot


Number of Plants
0 - 11
11 - 22
22 - 34
34 - 45
45 - 56
56 - 67
N
2 0 2 4 Km